5-2017

Novel statistical approaches for missing values in truncated high-dimensional metabolomics data with a detection threshold.

Jasmit SureshKumar Shah
University of Louisville

NOVEL STATISTICAL APPROACHES FOR MISSING VALUES IN TRUNCATED HIGH-DIMENSIONAL METABOLOMICS DATA WITH A DETECTION THRESHOLD

By

Jasmit SureshKumar Shah
B.S., University of South Alabama. 2009
M.S., University of Louisville, 2011

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Sciences
of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

May 2017

NOVEL STATISTICAL APPROACHES FOR MISSING VALUES IN TRUNCATED HIGH-DIMENSIONAL METABOLOMICS DATA WITH A DETECTION THRESHOLD

By

Jasmit SureshKumar Shah
B.S., University of South Alabama. 2009
M.S., University of Louisville, 2011

A Dissertation Approved on

April 14, 2017

by the following Dissertation Committee:

_____
Dissertation Director: Dr. Shesh N Rai

_____
Dissertation Co-Director: Dr. Guy N Brock

_____
Dr. Aruni Bhatnagar

_____
Dr. Jeremy Gaskins

_____
Dr. Dongfeng Wu

DEDICATION

In loving memory of my father, SureshKumar Lakhamshi Sumaria.

This Dissertation is dedicated to my mother Taruna SureshKumar Sumaria.

Thank you for your constant, unconditional love and support.

Mum, I love you!!!

ACKNOWLEDGEMENTS

Completion of my graduate career is a movement which is far from solitary and tough to pursue without the help and significant contributions from many wonderful people in my life. It is hard to capture my gratitude and appreciation in words, but each and every person acknowledged here have immensely contributed to my incredible journey both educationally and emotionally, and I thank them from the bottom of my heart to stand by with me.

I would never have discovered Biostatistics without the guidance of my undergraduate advisor Late Dr. Satya Mishra. He knew I was destined to be a statistician before I began my graduate studies, challenged me and guided me through opportunities that truly steered the course of my future.

Firstly, I would like to express my sincere gratitude to my mentors Drs. Shesh Rai and Guy Brock for their continuous support of my Ph.D. study and related research. Without their immense support, this degree would not have been possible.

Dr. Shesh Rai is an icon of guidance and leadership for his students and has always inspired never to give up. He places very high expectations on every student he supervises, and at the same time, he is caring and always looking for ways to improve the learning experiences of his students. He also exerts a substantial amount of energy into training his students and encouraging independent thinking and further assessment of the problem. With his support,

I also got an opportunity to work full time at the Diabetes and Obesity Center after my candidacy.

Dr. Guy Brock has been an excellent advisor, and I appreciate the wisdom, direction, and support he has given these past three years. He has always been a favorite among all the students, and I am glad I got to work with him for my dissertation. He has been a great mentor and has included me in his other projects as well, and has always been a tremendous help no matter the task or circumstance. His positive attitude and mentorship have allowed me to focus on my research area and has always assisted me in improving and capitalizing on essential practical skills with the Statistical background. I started the Ph.D. program with the hopes to have the Dr. Brock involved in my dissertation and I am glad to leave with a mentor, guide and friend.

With the guidance and motivation of my mentors, I have also had a chance to showcase my research at local and national conferences. The right structure provided to me and the answers to my endless questions has made me complete this dissertation, and I attribute much of my professional success to their guidance. Completing my Ph.D, I know my relation with Dr. Rai and Dr. Brock was not only for the past few years but for the many years ahead of me in my professional career.

I would like to thank my committee members, Dr. Aruni Bhatnagar, Dr. Jeremy Gaskins, and Dr. Dongfeng Wu, for their involvement and suggestions. My many thanks to Dr. Bhatnagar and all the individuals at the Diabetes and Obesity Center for being great colleagues at work. My involvement with collaborators at the Center has not only allowed me to contribute a lot in the applied medical research but also refined the statistical methods in the field. Many thanks to all the faculty members in the Department of Bioinformatics

and Biostatistics for their dedication to education and research. Their educational instruction significantly contributed to my academic growth and success.

Finally, I want to thank everyone who has been part of this incredible journey in the United States. There are two main things that I look up to every single day. One of them is someone I look up to, and the other is someone I look forward to. I want to thank God, who I look up to, who has beautified my life with occasions that I know are not of my hand. Appreciating what life has given me and focusing on all the positives has made me a better person and a firm believer that God is always guiding you whether we know it or not. To my family, who I look forward to, have stood by me thick and thin for me to fulfill my dreams. My father, whom I lost a year ago, always supported my dreams and ambitions, regardless of how unachievable they seemed. My mother, who always knew I would go high in my educational career and who has sacrificed far more than expected. She always wanted me to be a Doctor, and I am glad I could dedicate this Ph.D. to her. My brothers and sisters for sacrificing a lot, so that I could achieve this dream. They have been my biggest support, always have believed in me and encouraged me with focusing on my goals. Being the youngest sibling, they have always pampered me, and I am so grateful to have them in my life.

I would also like to thank Dr. Premhar Shah, who has always inspired me to dream big and reach my goals. He is not only a great uncle but a great person who has been a great inspiration. He always made sure I got the best of everything and had also visioned for me pursue high in my education career. He made certain that we got the best education in Eldoret, and today I can proudly say I am one of first Ph.D. graduates from my school in Eldoret. My thanks to Eldoret town and Gulab Lochab Academy, because that is where my

roots began. I am so grateful to have spent my seventeen years in Eldoret and my education at Gulab Lochab Academy, and to give them back a Ph.D. graduate is something I am very proud of.

Finally, I am very glad and proud of all my wonderful friends in the United States, especially in Louisville. Whether it is with ups and downs in my academic career or my personal life, they have been a great and excellent support. They have been very important in my life, and have made me grow to a person of who I am today.

"The harder you fall, the heavier your heart; the heavier your heart, the stronger you climb; the stronger you climb, the higher your pedestal" (Criss Jami)

ABSTRACT

NOVEL STATISTICAL APPROACHES FOR MISSING VALUES IN TRUNCATED HIGH-DIMENSIONAL METABOLOMICS DATA WITH A DETECTION THRESHOLD

Jasmit S Shah

April 14, 2017

Despite considerable advances in high throughput technology over the last decade, new challenges have emerged related to the analysis, interpretation, and integration of high-dimensional data. The arrival of omics datasets has contributed to the rapid improvement of systems biology, which seeks the understanding of complex biological systems. Metabolomics is an emerging omics field, where mass spectrometry technologies generate high dimensional datasets. As advances in this area are progressing, the need for better analysis methods to provide correct and adequate results are required. While in other omics sectors such as genomics or proteomics there has and continues to be critical understanding and concern in developing appropriate methods to handle missing values, handling of missing values in metabolomics has been an undervalued step.

Missing data are a common issue in all types of medical research and handling missing data has always been a challenge. Since many downstream analyses such as classification methods, clustering methods, and dimension reduction methods require complete datasets, imputation

of missing data is a critical and crucial step. The standard approach used is to remove features with one or more missing values or to substitute them with a value such as mean or half minimum substitution. One of the major issues from the missing data in metabolomics is due to a limit of detection, and thus sophisticated methods are needed to incorporate different origins of missingness.

This dissertation contributes to the knowledge of missing value imputation methods with three separate but related research projects. The first project consists of a novel missing value imputation method based on a modification of the *k* nearest neighbor method which accounts for truncation at the minimum value/limit of detection. The approach assumes that the data follows a truncated normal distribution with the truncation point at the detection limit. The aim of the second project arises from the limitation in the first project. While the novel approach is useful, estimation of the truncated mean and standard deviation is problematic in small sample sizes ($N < 10$). In this project, we develop a Bayesian model for imputing missing values with small sample sizes. The Bayesian paradigm has generally been utilized in the omics field as it exploits the data accessible from related components to acquire data to stabilize parameter estimation. The third project is based on the motivation to determine the impact of missing value imputation on down-stream analyses and whether ranking of imputation methods correlates well with the biological implications of the imputation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

The advent of high-throughput technology to generate massive datasets in biomedical research has been on a high rise. Successively, new challenges emerged related to the analysis, interpretation, and integration of such data. The diversity of technological advances drives the need for efficient analytical methods. Developments in biomedical research within molecular biology now allow simultaneous measurements of thousands of cellular components at different hierarchical levels, such as genomics, proteomics, transcriptomics, and metabolomics. In the "-omics" field, metabolomics is a growing field showing great potential in identifying relevant metabolites in biomedical research. It involves the biochemical profiling of all the metabolites in a cell, tissue, or organism, and focuses on the best measurement of the physiological state of organism's metabolites (Schmidt, 2004). There has been substantial progress in the development of high-throughput methods for metabolomics in the last decade with rapid improvements in mass spectrometry (MS)-based methods (Shah et al., 2000), and in computer hardware and software that is adept at handling large datasets (Katajamaa and Oresic, 2007). A wide range of mass spectrometric techniques have been used in metabolomics and the most popular methods used are GC-MS (gas chromatography mass spectrometry), LC-MS (liquid chromatography-mass spectrometry) and NMR (nuclear magnetic resonance). In spite of the fact that metabolomics has the likelihood of providing understanding to numerous biological questions, data generated by mass spectrometry pose some statistical challenges.

Missing values (MV) are challenging since most statistical analyses require a complete dataset. They can occur for several different reasons including equipment malfunction, sample contamination, and sporadic missed measurements. Many of the studies will have more than one type of MV and appropriately handling MVs is important in the inference for a parameter. Based on different statistical techniques, MVs are dealt with differently. A common approach is to use the complete case analysis method, i.e. removing cases with missing values for any of the variables. The other standard approach is done by filling in (imputing) plausible values for the MVs, making more efficient use of the data. Often a study will have more than one type of MV, although they are treated as the same kind. Treating each of these types of MVs separately has its advantage. One advantage of imputing one type of MV first computationally simplifies the imputation of the rest of the MVs. Another advantage is that treating the MVs differently allows the researcher to compute how much variability and how much missing information is due to each type of MV. Studies which generate high dimensional data can have MVs of all three types, categorized as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Details in Missing Values section below) and to handle MVs with all the types has been an unexplored area of research.

Metabolomics

The 'omics field has become a popular and hot research area in biomedical studies due to its detailed content of the cells, tissues, organs or biofluids provided by high throughput technologies. Metabolomics is a relatively new area in the omics field, with the term "metabolome," devised less than two decades ago (Oliver et al. 1998). It is the study of small molecules (molecular weight < 1,000 Da) in a biological system using high-throughput identification and quantification techniques. These molecules, measured simultaneously provide an insight into the functioning of metabolic pathways of the whole biological system

for its selected cellular, tissue or organ levels (Fiehn, 2002). Within the omics cascade, metabolomics is further down in line from genomics, proteomics, and transcriptomics and is believed to easily correlate with the phenotype as the metabolites serve as direct signatures of biochemical activity. Metabolomics comprises the qualitative and quantitative analysis of metabolites using two approaches: targeted and untargeted metabolomic analysis. Targeted metabolomic analysis focuses on quantitative changes in metabolites of interest (e.g., amino acids, carbohydrates, steroids, and fatty acids) based on a priori knowledge of the biological function or metabolic pathway whereas untargeted metabolomic analysis involves the identification and characterization of a vast number of metabolites and their precursors. (Sadanala et al., 2012). The two most relevant technical approaches for the generation of metabolomic datasets are mass spectrometry (MS) and nuclear magnetic resonance (NMR). MS is an analytical method that obtains spectral data in the form of a mass-to-charge ratio (m/z) and a relative intensity of the measured compounds (Alonso 2015). The biological sample first needs to be ionized for the peak signals to be generated for each metabolite. The ionized compounds from each molecule will then produce different peak patterns that define the impression of the original molecule. Before the MS quantification the separation step is performed, where the complexity of the biological sample is reduced to allow the MS analysis of different sets of molecules at various times. The most common separation methods used are liquid and gas chromatography (LC and GC, respectively) (Theodoridis et al., 2011). The LC or GC separation techniques is based on the interaction of the different metabolites in the sample with the adsorbent materials used in the chromatography, and thus this way, molecules with different chemical properties will require different amounts of time to pass through the column. NMR is the other primary approach used based on spectroscopic technique. It relies on the energy absorption, and re-emission of the atom nuclei due to variations in an external

magnetic field (Alonso 2015, Bothwell and Griffin 2011). The quantification of the concentrations of molecules are based on the spectral data from NMR, which also provides information about its chemical structure. The spectral peak areas generated by each molecule are used as an indirect measure of the quantity of the metabolite in the sample, while the pattern of spectral peaks informing on the physical properties of the molecule is used to identify the type of metabolite (Alonso 2015).

In Figure 1, the conventional pipeline for generating a metabolic profile is shown. The typical workflow that is commonly used in high-throughput metabolomic studies starts with the processing of the biological samples to produce the metabolic information. Different techniques mentioned such as LC-MS, GC-MS or NMR is used for the spectral identification and are then processed by various methods. A detailed pre-processing of the spectra data is performed using baseline correction, noise reduction, smoothing, peak detection and alignment and peak integration. Once the complete set of the metabolic profile has been generated, data analysis methods such as univariate and multivariate analyses can be applied to study the general structure of the metabolomics data and how different metabolites are related to the phenotypic data associated with the samples.

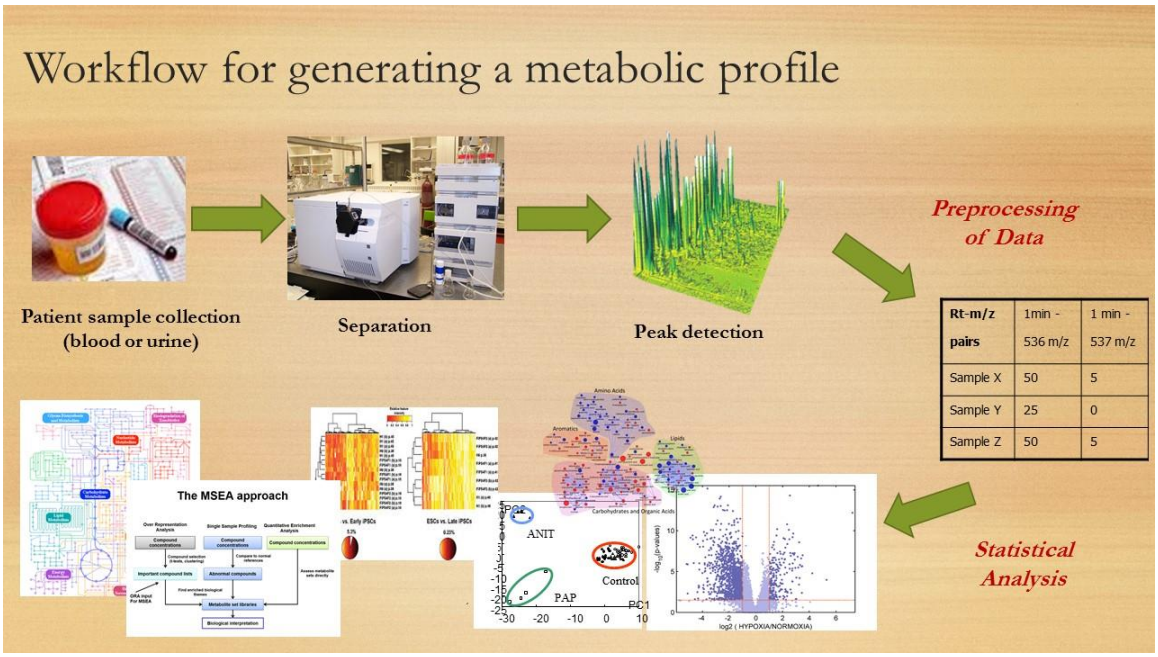

Figure 1: The analysis workflow in generating a metabolic profile and the various steps of the metabolomic analysis pipeline

Missing Values

MVs are the unobserved values in a data set which can be of various types and may be missing for different reasons. The various reasons why missingness occurs could be due to human error, equipment malfunction, dropouts, latent variables. MVs in metabolomics generally arise due to a number of reasons, such as: (1) limits in computational detection; (2) imperfection of the algorithms whereby they fail in the identification of some of the signals from the background; (3) low intensity of the signals used; (4) measurement error; and (5) deconvolution that may result in false negative during separation of overlapping signals (Gromski et al 2014). MVs can be problematic across many fields, and appropriate methods typically need to be considered when analyzing incomplete data. Knowledge about the nature of the missing values can help identify the most appropriate method for dealing with missing data (Little & Rubin, 2002). MVs are categorized based on a mechanism where it describes the relationship between the probability of a value being missing and the other variables in the dataset. Let Y represent the complete dataset that can be seperated as (Y_{obs}, Y_{miss}) where Y_{obs} is the observed values and Y_{miss} is the missing values. Let R be an indicator variable indicating whether a value is observed or missing, where $R = 1$ denotes a value which is observed and $R = 0$ denotes a value which is missing. The matrix R stores the location of the MVs and its distribution may depend on $Y = (Y_{obs}, Y_{miss})$, either by design or by coincidence and this relation is described by the missing data model, $\Pr(R|Y = (Y_{obs}, Y_{miss}), \varphi)$, where φ contains the parameters of the missing data model. The following are the three mechanisms of missingness as described in Rubin (1976) and Rubin (1987).

The first mechanism of missingness is missing at random (MAR). If the probability of missing is the same within groups defined by the observed data, then the data are MAR. This mechanism of missingness is given by

$$\Pr(R = 0|Y, \varphi) = \Pr(R = 0|Y_{obs}, \varphi)$$

That is the probability of missingness is only dependent on the observed data and not the unobserved/missing data. A simple example of MAR is a depression survey where male subjects are more likely to refuse to fill out the survey, although it does not depend on the level of their depression.

The second mechanism of missingness is missing completely at random (MCAR). If the probability of being missing is the same for all cases, then the data are MCAR. This mechanism of missingness is given by

$$\Pr(R|Y, \varphi) = \Pr(R| \varphi)$$

That is the probability of missingness is not dependent on the observed data and the unobserved/missing data. A simple example of MCAR is if a set of household income values are missing and if the percentages of missing are equal among ethnicity group, gender, and educational group, then the missingness is MCAR, a special case of MAR.

The third mechanism of missingness is missing not at random (MNAR). If neither MCAR nor MAR holds, then it is MNAR, where the probability of being missing is dependent on the observed and the unobserved/missing data.

$$\Pr(R = 0|Y, \varphi) = \Pr(R = 0|Y_{obs}, Y_{miss}, \varphi)$$

One example of MNAR is where subjects with severe depression or side effects from the medication are more likely to be missing at the end of the study.

Downstream analyses via multivariate methods require a complete dataset. MVs are handled differently and thus affects the interpretation and statistical inference. One common approach used is case deletion or complete case analysis, wherein this method only completed cases with no MVs are included in the analysis. Case deletion leads to a smaller sample size and several articles show examples using case removal and results with low power (Harel et al., 2012; White & Carlin, 2010). Another approach to handling MVs is via single imputation methods where MVs are filled in with plausible values. It is a widely used method and is pretty straight forward but also a dangerous way of dealing with missing values. Statistical analysis performed on datasets imputed by single imputation method may be biased as the approach does not consider the uncertainty of the imputed values. Some of the single imputation methods include mean, zero, half minimum and median imputation where the MVs are replaced by the mean, zeros, half of the minimum and median of the variable respectively. The magnitude of the covariances and correlation also decreases by limiting the variability, and this method often causes biased estimates, irrespective of the underlying missing data mechanism (Enders, 2010; Eekhout et al., 2012). Other methods in single imputation are based on computation such as imputation using k-nearest neighbor (kNN), random forest (RF), Bayesian principal component analysis (BPCA), probabilistic principal component analysis (PPCA), and singular value decomposition (SVD) imputation. Many of the single imputation methods are thoroughly described in Schafer and Graham (2002). Other sophisticated methods of dealing with MVs include multiple imputation (Rubin 1987), nonparametric imputation (Wang&Chen 2009), hot deck imputation (Andridge & Little 2010), weighting techniques (Meng, 1994), maximum likelihood (Little and Rubin 2002) via the EM algorithm (Dempster et al 1997) and Bayesian analysis (Gelman et al 2003). Most of these methods assume the data is MAR or MCAR and none of the methods combine the MNAR mechanism directly. There is noticeable

absence in the literature of imputation methods that account for MAR and MNAR mechanisms and thus the motivation to develop a method that accounts for both the mechanisms.

Dissertation Outline

In this dissertation, we develop two novel approaches for imputing MVs which can simultaneously handle missing data generated by both MNAR and MAR mechanisms. We further investigate the impact of data imputation on statistical analyses. The rest of the dissertation is organized as follows. In Chapter 2, we develop a novel missing value imputation method based on a modification of the k nearest neighbor method which accounts for truncation at the minimum value/limit of detection. The approach assumes that the data follows a truncated normal distribution with the truncation point at the detection limit. In Chapter 3, we develop a Bayesian model for imputing missing values with small sample sizes. The aim of the project arises from the limitation in the previous project. While the novel approach is useful, estimation of the truncated mean and standard deviation is problematic in small sample sizes ($N < 10$). The Bayesian paradigm has generally been utilized in the omics field as it exploits the data accessible from related components to acquire data to stabilize parameter estimation. In Chapter 4, we investigate a comprehensive analysis on the impact of missing value imputation on down-stream analyses focusing on differentially expressed metabolite detection, classification and clustering analyses. Finally, in Chapter 5 we finish with some concluding remarks and potential future research.

CHAPTER 2[1]

TRUNCATION BASED NEAREST NEIGHBOR IMPUTATION FOR HIGH DIMENSIONAL DATA WITH DETECTION LIMIT THRESHOLD

2.1 Background

High throughput technology makes it possible to generate high dimensional data in many areas of biochemical research. Mass spectrometry (MS) is one of the important high-throughput analytical techniques used for profiling small molecular compounds, such as metabolites, in biological samples. Raw data from a metabolomics experiment usually consist of the retention time (if liquid or gas chromatography is used for separation), the observed mass to charge ratio, and a measure of ion intensity (Taylor, Leiserowitz et al. 2013). The ion intensity represents the measure of each metabolite's relative abundance whereas the mass-to-charge ratios and the retention times assist in identifying unique metabolites. A detailed pre-processing of the raw data, including baseline correction, noise reduction, smoothing, peak detection and alignment and peak integration, is necessary before analysis (Want and Masson 2011). The end product of this processing step is a data matrix consisting of the unique features and its intensity measures in each sample. Commonly, data generated from MS have many missing values. Missing values (MVs) in MS can occur from various sources both technical and biological. There are three common sources of missingness: (Taylor, Leiserowitz et al. 2013) i) a metabolite could be truly missing from a sample due to biological reasons, ii) a

[1] The text and figures of this chapter were published in BMC Bioinformatics. 2017 Feb 20. 18:114. doi: 10.1186/s12859-017-1547-6

metabolite can be present in a sample but at a concentration below the detection limit of the MS, and iii) a metabolite can be present in a sample at a level above the detection limit but fail to be detected due to technical issues related to sample processing.

The limit of detection (LOD) is the smallest sample quantity that yields a signal that can be differentiated from the background noise. Shrivastava et al (Shrivastava and Gupta 2011) give different guidelines for the detection limit and describe different methods for calculating the detection limit. Some common methods (Shrivastava and Gupta 2011) for the estimation of detection limits are visual definition, calculation from signal to noise ratio, calculation from standard deviation (SD) of the blanks and calculation from the calibration line at low concentrations. Armbruster et al (Armbruster, Tillman et al. 1994) compare the empirical and statistical methods based on gas chromatography MS assays for drugs. They explain the calculation from SD where a series of blank (negative) samples (a sample containing no analyte but with a matrix identical to that of the average sample analyzed) are tested and the mean blank value and the SD are calculated, where the LOD is the mean blank value plus 2 or 3 SDs (Armbruster, Tillman et al. 1994). The signal-to-noise ratio method is commonly applied to analytical methods that exhibit baseline noise (Shrivastava and Gupta 2011, Cole, Mills et al. 2016). In this method, the peak-to-peak noise around the analyte retention time is measured, and subsequently, the concentration of the analyte that would yield a signal equal to a signal-to-noise ratio (S/N) of three is generally accepted for estimating the LOD (Shrivastava and Gupta 2011).

Missing data can be classified into three categories based on the properties of the causality of the missingness (Little and Rubin 2002): "missing completely at random (MCAR)", "missing at random (MAR)" and "missing not at random (MNAR)". The missing values are considered

MCAR if the probability of an observation being missing does not depend on observed or unobserved measurements. If the probability of an observation being missing depends only on observed measurements then the values are considered as MAR. MNAR is when the probability of an observation being missing depends on unobserved measurements. In metabolomics studies, we assume that the missing values occurs either as MNAR (metabolites occur at low abundances, below the detection limit) or MAR, e.g. metabolites are truly not present or are above the detection limit but missing due to technical errors. The majority of imputation algorithms for high-throughput data exploit the MAR mechanism and use observed values from other genes / proteins / metabolites to impute the MVs. However, imputation for MNAR values is fraught with difficulty (Karpievitch, Dabney et al. 2012, Taylor, Leiserowitz et al. 2013). Using the imputation methods for microarray studies in MS omics studies could lead to biased results because most of the imputation techniques produce unbiased results only if the missing data are MCAR or MAR (Karpievitch, Stanley et al. 2009). Karpievitch et al (Karpievitch, Dabney et al. 2012) discuss several approaches in dealing with missing values, considering MNAR as censored in proteomic studies.

Many statistical analyses require a complete dataset and therefore missing values are commonly substituted with a reliable value. Many MV imputation methods have been developed in the literature in other -omic studies. For example the significance of appropriate handling of MVs has been acknowledged in the analysis of DNA microarray (Troyanskaya, Cantor et al. 2001) and gel based proteomics data (Pedreschi, Hertog et al. 2008, Albrecht, Kniemeyer et al. 2010). Brock et al (Brock, Shaffer et al. 2008) evaluated a variety of imputation algorithms with expression data such as KNN, singular value decomposition, partial least squares, Bayesian principal component analysis, local least squares and least squares adaptive. In MS data analysis, a common approach is to drop individual metabolites with a large proportion of

subjects with missing values from the analysis or to drop the entire subject with a large number of missing metabolites. Other standard methods of substitution include using a minimum value, mean, or median value. Gromski et al (Gromski, Xu et al. 2014) analyzed different MV imputation methods and their influence on multivariate analysis. The choice of imputation method can significantly affect the results and interpretation of analyses of metabolomics data (Hrydziuszko and Viant 2011).

Since missingness may be due to a metabolite being below the detection limit of the mass spectrometer (MNAR) or other technical issues unrelated to the abundance of the metabolite (MAR), we develop a method that accounts for both of these mechanisms. To demonstrate missing patterns, Figure 2 summarizes the distribution of two different metabolites taken from Sansbury et al (Sansbury, DeMartino et al. 2014), both of which had missing values. The top graph shows that the distribution of the metabolite is far above the detection limit and therefore replacing the MV in that metabolite with a LOD value would be inappropriate. Similarly, the bottom graph shows that the distribution of the metabolite is near the detection limit and therefore replacing the MV with a mean or median value might be inappropriate.

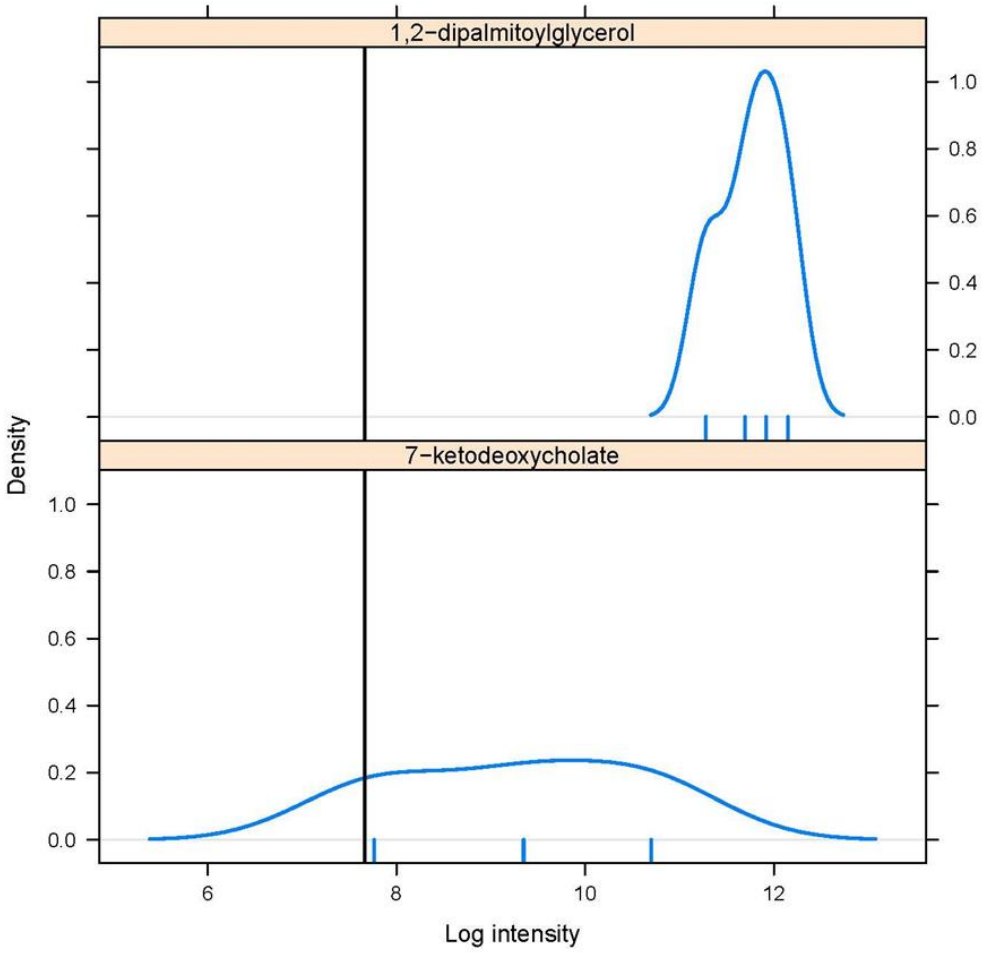

Figure 2. Two examples of metabolite distributions which have missing values (MVs), from the myocardial infarction data (Sansbury, DeMartino et al. 2014).

The black vertical line on each graph shows the minimum value of the data, considered as the lower limit of detection (LOD). The small vertical lines below the x-axis in each case indicate the observed values of the metabolites. The figure on the top shows the distribution of 1,2 dipalmitoylglycerol, where the observed values are all around 3 standard deviations above the LOD. In this case, the MVs are likely to be MAR or MCAR. In contrast, the figure on the bottom shows the distribution of 7-ketodeoxycholate, which is close to the LOD. Here, the MVs are likely to be below the LOD and hence MNAR.

In this work, we develop an imputation algorithm based on nearest neighbors that considers MNAR and MAR together based on a truncated distribution, with the detection limit considered as the truncation point. The proposed truncation-based KNN method is compared to standard KNN imputation based on Euclidean and correlation based distance metrics. We show that this method is effective and generally outperforms the other two KNN procedures through extensive simulation studies and application to three real data sets (Sansbury, DeMartino et al. 2014, DeFilippis, Chernyavskiy et al. 2016).

2.2 Methods

K-Nearest Neighbors (NN)

KNN is a non-parametric machine learning algorithm. NN imputation approaches are neighbor based methods where the imputed value is either a value that was measured for the neighbor or the average of measured values for multiple neighbors. It is a very simple and powerful method. The motivation behind the NN algorithm is that samples with similar features have similar output values. The algorithm works on the premise that the imputation of the unknown samples can be done by relating the unknown to the known according to some distance or similarity function. Essentially, two vectors that are far apart based on the distance function are less likely than two closely situated vectors to have a similar output value. The most frequently used distance metrics are the Euclidean distance metric or the Pearson correlation metric. Let $X_i, i = 1, \dots, n$ be independent and identically distributed (iid) with mean μ_X and standard deviation σ_X, and $Y_i, i = 1, \dots, n$ be iid with mean μ_Y and standard deviation σ_Y. The two sets of measurements are assumed to be taken on the same set of observations. Then the Euclidean distance between the two sample vectors $\boldsymbol{x} = \langle x_1, \dots, x_n \rangle$ and $\boldsymbol{y} = \langle y_1, \dots, y_n \rangle$ is defined as follows:

$$d_E(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

It is the ordinary distance between two points in the Euclidean space. The correlation between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as follows:

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{\frac{1}{n}\sum_i x_i y_i - \hat{\mu}_X \hat{\mu}_Y}{\hat{\sigma}_X \hat{\sigma}_Y}$$

where $\hat{\mu}_X$, $\hat{\mu}_Y$, $\hat{\sigma}_X$, and $\hat{\sigma}_Y$ are the sample estimates of the corresponding population parameters. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are standardized (denoted as $\boldsymbol{x}^s$ and $\boldsymbol{y}^s$, respectively) to each have a mean of zero and a standard deviation of one, the formula reduces to:

$$r(\boldsymbol{x}^s, \boldsymbol{y}^s) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

When using the Euclidean distance, normalization/re-scaling process is not required for KNN imputation because neighbors with similar magnitude to the metabolite with MV are used for imputation. In the correlation based distance, since metabolites can be highly correlated but different in magnitude, the metabolites are first standardized to mean zero and standard deviation one before the neighbor selection and then re-scaled back to the original scale after imputation (Brock, Shaffer et al. 2008, Tutz and Ramzan 2015) . The distance used to select the neighbors is $d_C = 1 - |r|$, where r is the Pearson correlation. This distance allows for information to be incorporated from both positively correlated and negatively correlated neighbors. During the distance calculation MVs are omitted, so that it is based only on the complete pairwise observations between two metabolites.

The KNN based on the Euclidean (KNN-EU) or Correlation (KNN-CR) distance metrics do not account for the truncation at the minimum value or the limit of detection. In our method, we propose a modified version of the KNN approach which accounts for the truncation at the minimum value called KNN Truncation (KNN-TN). A truncated distribution occurs when there is no ability to know about data that falls below a set threshold or outside a certain set range. Often the general idea is to make inference back to the original population and not on the truncated population and therefore inference is made on the population mean and not the truncated sample mean. In the regular KNN-CR, the metabolites are standardized based on the sample mean and sample standard deviation. In KNN-TN, we first estimate the means and standard deviation, and use the estimated values for standardizing. Maximum likelihood Estimators (MLE) are estimated for the truncated normal distribution. The likelihood for the truncated normal distribution is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{P(Y \in (a, \infty) | \mu, \sigma^2)} \right) \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) e^{\frac{-(y_i - \mu)^2}{2\sigma^2}}$$

Here *a* is the truncation point and presumed to be known in our case. Also note that MVs are ignored and the likelihood is based only on the observed data (in essence a partial likelihood akin to a Cox regression model (Efron 1977, Ren and Zhou 2010). The log likelihood is then

$$\begin{aligned} l &= \ln L(\mu, \sigma^2) \\ &= -n \ln\left(P(Y \in (a, \infty) | \mu, \sigma^2)\right) - n \ln\left(\sqrt{2\pi\sigma^2}\right) - \frac{\sum (y_i - \mu)^2}{2\sigma^2} \end{aligned}$$

The $P(Y \in (a, \infty) | \mu, \sigma^2)$ is the part of the likelihood that is specific to the truncated normal distribution.

We use the Newton-Raphson (NR) optimization procedure to find the MLEs for μ and σ (Cohen 1949, Cohen 1950) (for details see the Appendix 1). The sample means and standard deviations are used as the initial values for the NR optimization. To accelerate the run-time of the algorithm, truncation-based estimation of the mean and standard deviation was done only on metabolites that had a sample mean within 3 standard deviations of the LOD. For the other metabolites, we simply used the sample means and standard deviations. The runtime for one dataset with 50 samples and 400 metabolites and the three missing mechanisms was about 1.20 minutes on average, which included truncation-based estimation of the mean and standard deviation and the three imputation methods. In particular for one individual run on 50 samples and 400 metabolites with 15% missingness, the runtime was about 1.81 seconds for the KNN-EU method, 3.41 seconds for the KNN-CR method and 19.95 seconds for the KNN-TN method. The KNN-TN method runtime was a little longer due to the estimation of the means and standard deviations.

Let y_{im} be the intensity of metabolite $m\ (1 \leq m \leq M)$ in sample $i\ (1 \leq i \leq N)$. The following steps outline the KNN imputation algorithms (KNN-TN, KNN-CR, and KNN-EU):

1. Choose a *K* to use for the number of nearest neighbors.
2. Select the distance metric: Euclidean (KNN-EU) or correlation (KNN-CR and KNN-TN)
3. If using correlation metric, decide whether to standardize the data based on sample mean and sample standard deviation (KNN-CR) or the truncation-based estimate of the mean and standard deviation (KNN-TN).

4. Based on the distance metric and (possibly) standardization, for each metabolite with a missing value in sample *i* find the *K* closest neighboring metabolites which have an observed value in sample *i*.
5. For metabolite *m* with missing value in sample *i*, calculated the imputed value $\hat{y}_{im}$ by taking the weighted average of the *K* nearest neighbors for each missing value in the metabolite. The weights are calculated as $w_k = \text{sign}(r_k)\, d_k^{-1} / \sum_{l=1}^{K} d_l^{-1}$, where $d_1, \ldots, d_K$ are the distances between metabolite *m* and each of the *K* neighbors and $r_1, \ldots, r_K$ are the corresponding Pearson correlations. The multiplication by $\text{sign}(r_k)$ allows for incorporation of negatively correlated metabolites. The imputed value is then $\hat{y}_{im} = \frac{1}{K} \sum_{k=1}^{K} w_k y_{ik}$.
6. If using the KNN-CR or KNN-TN approaches, back-transform into the original space of the metabolites.

The steps for the KNN-TN procedure are outlined graphically in Figure 3 (see figure caption for detailed explanation). The graph illustrates the algorithms success at imputing both MAR and MNAR values.

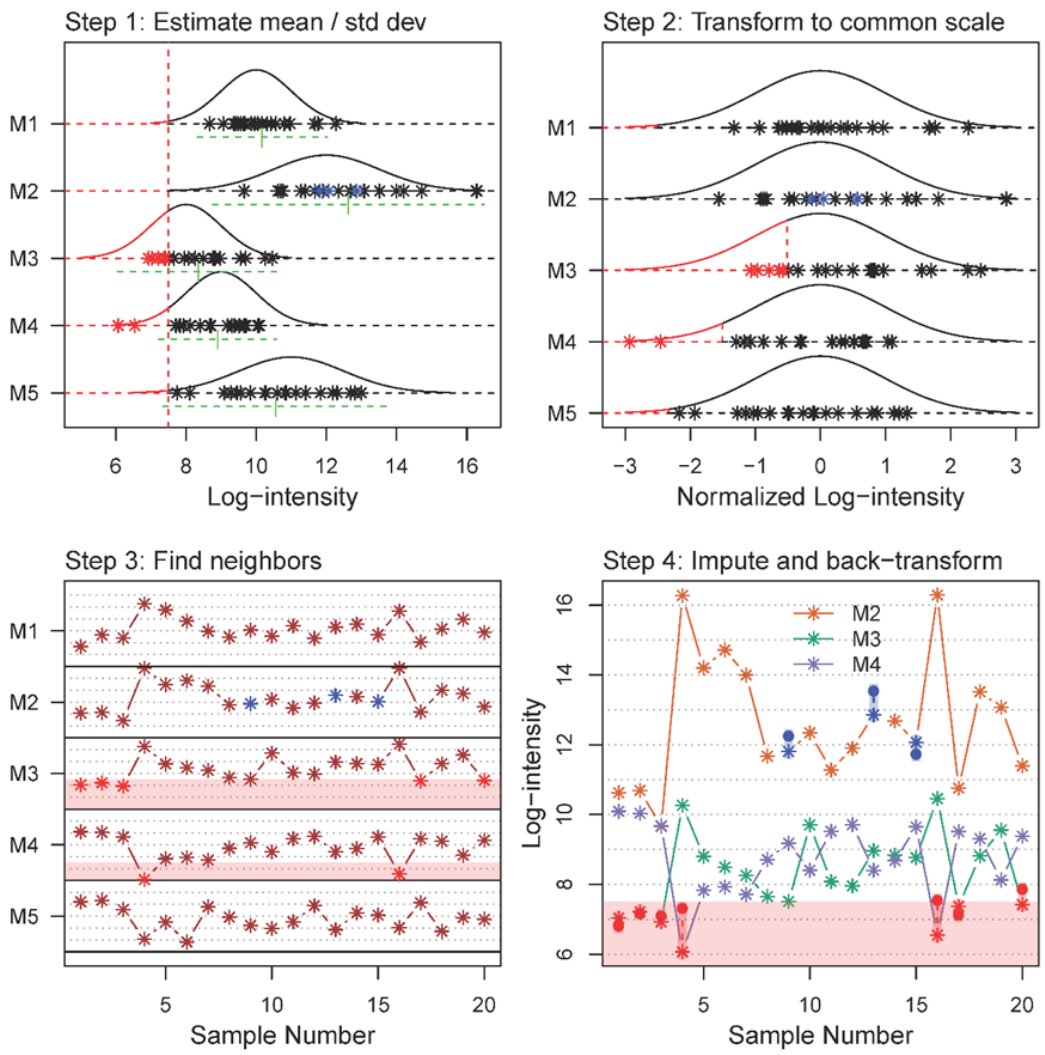

Figure 3: Steps in the KNN-TN imputation algorithm.

Step 1 (top left panel): The first step in the KNN-TN procedure is to estimate the mean and standard deviation of each metabolite. Here, the distribution and simulated values for 5 metabolites (M1-M5) and 20 samples are given. For each metabolite, observed values are given by black stars. Additionally, M2 has 3 values that are MAR (blue stars), while M3 has 5 points that are MNAR (below the LOD, red stars) and M4 has 2 points below the LOD (red stars). The estimate mean for each metabolite is indicated by the underlying green vertical

dash, while the green horizontal dashed line represents the estimated standard deviation (the line extends out +/- 2 standard deviations).

Step 2 (top right panel): The second step in the procedure is to transform all the values to a common scale, with mean zero and standard deviation of one for each metabolite. The original points are represented in this transformed scale with black stars, with MNAR values in red and MAR values in blue.

Step 3 (bottom left panel): The next step is to find metabolites with a similar profile on this common scale. In this case, metabolites M1-M3 are highly correlated and M4-M5 are also highly correlated. The two groups of metabolites are also negatively correlated with each other, and this information can also be used to aid the imputation process. The missing values are imputed in the transformed space, with weights based on the inverse of the distances $1 - |r|$ (r is the Pearson correlation between the two metabolites). Contributions from negatively correlated metabolites are multiplied by negative one. The region below the LOD is shaded light red.

Step 4 (bottom right panel): The values are then back-transformed to the original space based on the estimated means and standard deviations from Step 1. Here, we show the three metabolites with missing values M2, M3, and M4. Solid circles represent imputed values for MAR (blue circles) and MNAR (red circles). The region below the LOD is again shaded in light red, while the slightly darker shaded regions connect the imputed value with its underlying true value. The imputed values are fairly close to the true values for metabolites M2 and M3, while for metabolite M4 the values are further away due to under-estimation of the true variance for M4 (c.f. top left panel).

Assessment of Performance

We evaluated the performance of the imputation methods by using the root mean squared error (RMSE) as the metric. It measures the difference between the estimated values and the original true values, when the original true values are known. The following simulation procedure from a complete dataset with no MVs is performed. MVs are generated by removing a proportion p of values from the complete data to generate data with MVs. The MVs are then imputed as $\hat{y}_{im}$ using the given imputation method. Finally, the root mean squared error (RMSE) is used to assess the performance by comparing the values of the imputed entries with the true values:

$$RMSE = \sqrt{\frac{1}{n(\mathcal{M})}\sum_{y_{im} \in \mathcal{M}}(\hat{y}_{im} - y_{im})^2},$$

where $\mathcal{M}$ is the set of missing values and $n(\mathcal{M})$ is the cardinality or number of elements in $\mathcal{M}$. Statistical significance of differences in RMSE values between methods was determined using multi-factor ANOVA models (with pre-defined contrasts for differences between the methods), with main effects for each factor in the simulation study. We further evaluate the biological impact of MV imputation on downstream analysis, specifically analyzing differences in mean log intensity between groups via the t-test. We evaluate the performances of the MV imputation using the metabolite list concordance index (MLCI) (Oh, Kang et al. 2010). By applying a selected MV imputation method, one metabolite list is obtained from the complete data and another is obtained from the imputed data. The MLCI is defined as:

$$MLCI\ (M_{CD}, M_{ID}) = \frac{n(M_{CD} \cap M_{ID})}{n(M_{CD})} + \frac{n(M_{CD}^C \cap M_{ID}^C)}{n(M_{CD}^C)} - 1\ ,$$

where M_{CD} is the list of statistically significant metabolites in the complete data, M_{ID} is the list of statistically significant metabolites in the imputed data, and M_{CD}^{C} and M_{ID}^{C} represent their complements, respectively. The metabolite list taken from the complete dataset is considered as the gold standard and a high value in MLCI indicates that the metabolite list from the imputed data is similar to that from the complete data.

2.3 Simulation Studies

We carried out a simulation study to compare the performance of the three different KNN based imputation methods. The simulations were conducted with 100 replications and are similar in spirit to those used in Tutz and Ramzan, 2015 (Tutz and Ramzan 2015) . For each replication we generated data with different combinations of sample sizes n and number of metabolites m. Each set of metabolites for a given sample were drawn from a m dimensional multivariate normal distribution with a mean vector μ and a correlation matrix Σ. We consider, in particular, three structures of the correlation matrix: blockwise positive correlation, autoregressive (AR) type correlation and blockwise mixed correlation.

Blockwise correlation

Let the columns of the data matrix $Y_{(N \text{ X } M)}$, be divided into B blocks, where each block contains M/B metabolites. The partitioned correlation matrix has the form

$$\sum = \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1B} \\ \vdots & \ddots & \vdots \\ \Sigma_{B1} & \cdots & \Sigma_{BB} \end{pmatrix}$$

The matrices Σ_{ii} are determined by the pairwise correlations ρ_w, such that all the components have a within block correlation of ρ_w. The matrices $\Sigma_{ij}, i \neq j$, are determined by the pairwise correlations ρ_{off} ; that is, all the components have a between block correlation ρ_{off}. The two

types of blockwise correlation matrices used in this study are one with all positive correlations where the ρ_w is positive only and the other is mixed where Σ_{ii} contains both positive and negative correlations. The mixed correlation has the form which is blockwise split in half where the diagonal blocks are positively correlated and the off-diagonal blocks are negatively correlated. For example, if Σ_{ii} contained 6 metabolites for any i, the matrix Σ_{ii}would be:

$$\sum_{ii} = \begin{pmatrix} 1 & + & + & - & - & - \\ + & 1 & + & - & - & - \\ + & + & 1 & - & - & - \\ - & - & - & 1 & + & + \\ - & - & - & + & 1 & + \\ - & - & - & + & + & 1 \end{pmatrix}$$

where the $+$ is the positive ρ_w and $-$ is the negative ρ_w

Autoregressive-type correlation

The other correlation structure used is the autoregressive type correlation. An AR correlation matrix of order one is defined by pairwise correlations $\rho^{|i-j|}$, for metabolites$i, j = 1, \dots, M$.

The combinations used were (N [Samples] X M [Metabolites]) = 20 X 400, 50 X 400, and 100 X 900. The means of the metabolites are assumed to be different and are generated from a $\text{Uniform}(-5,5)$ distribution. The metabolites within each block were strongly correlated with $\rho_w = 0.7$, but nearly uncorrelated with metabolites in other blocks, $\rho_{\text{off}} = 0.2$. In the AR type correlation ρ = 0.9. For the degree of missing, three levels were studied: 9% missing, 15% missing and 30% missing. Missing data were created based on the two kinds of missingness, MNAR and MAR (technically the latter are generated by MCAR, though a MAR mechanism can be exploited for imputation since the metabolite values are highly correlated). Within each level of missing, a one-third and two-third combination was used to create both MNAR and MAR. We looked at the scenario where MNAR is greater than MAR

and vice versa. For example in 9% missing, we considered 6% as MNAR and 3% as MAR and then considered 6% as MAR and 3% as MNAR. Data below the given MNAR percent was considered as missing and the MAR percent was randomly generated in the non-missing data. The datasets with missing values were passed through a cleaning process where metabolites with more than 75% missing observations were eliminated individually. Throughout, the number of neighbors *K* used for imputation was set to 10. We evaluated three *K*'s (*K*=5, 10 and 20) and found consistency in *K*=10 as it gave the best RMSE values.

2.4 Real Data Studies

Myocardial Infarction Data

We used the in vivo metabolomics data on myocardial infarction (MI). The data consists of two groups, MI vs control, 5 samples in each group and 288 metabolites. Adult mice were subjected to permanent coronary occlusion (myocardial infarction; MI) or Sham surgery. Adult C57BL/6J mice from The Jackson Laboratory (Bar Harbor, ME) were used in this study and were anesthetized with ketamine (50 mg/kg, intra-peritoneal) and pentobarbital (50 mg/kg, intra-peritoneal), orally intubated with polyethylene-60 tubing, and ventilated (Harvard Apparatus Rodent Ventilator, model 845) with oxygen supplementation prior to the myocardial infarction. The study was aimed to examine the metabolic changes that occur in the heart in vivo during heart failure using mouse models of permanent coronary ligation. A combination of liquid chromatography (LC) MS/MS and gas chromatography (GC) MS techniques was used to measure the 288 metabolites in these hearts. The MS was based on a Waters ACQUITY UPLC and a Thermo-Finnigan LTQ mass spectrometer, which consisted of an electrospray ionization source and linear ion trap mass analyzer. The cases had 220 metabolites with complete values, 6 metabolites with complete missing and 62 metabolites had

4.8% missing values whereas the controls had 241 metabolites with complete values, 7 metabolites with complete missing and 40 metabolites had 7.8% missing values. The LOD for this dataset is considered as the minimum value of the dataset as commonly used in untargeted metabolomics. Details of the experiments are described in Sansbury et al (Sansbury, DeMartino et al. 2014).

Atherothrombotic Data

We used the human atherothrombotic myocardial infarction (MI) metabolomics data. The data was identified between two groups, those with acute MI and those with stable coronary artery disease (CAD). Acute MI was further stratified into thrombotic (Type1) and non-thrombotic (Type2) MI. The data was collected across four time points and for the context of this research we used the baseline data only. The three groups, sCAD, Type1 and Type2 had 15, 11, and 12 patients with 1032 metabolites. The sCAD had 685 metabolites with complete values, 39 metabolites with complete missing, and 308 metabolites had 10.2% missing, the Type1 group had 689 metabolites with complete values, 43 metabolites with complete missing and 300 metabolites had 9.8% missing whereas the Type2 group had 610 metabolites with complete values, 66 metabolites with complete missing and 356 metabolites had 12.3% missing. The LOD for this dataset is considered as the minimum value of the dataset as commonly used in untargeted metabolomics. Plasma samples collected from the patients were used and 1032 metabolites were detected and quantified by GC-MS and ultra-performance (UP) LC-MS in both positive and negative ionization modes. Details of the experiment are described in DeFilippis et al (DeFilippis, Chernyavskiy et al. 2016).

African Race Data

We used the African Studies data which is publicly available on The Metabolomics WorkBench. This data is available at the NIH Common Fund's Data Repository and Coordinating Center (supported by NIH grant, U01-DK097430) website (http://www.metabolomicsworkbench.org), where it has been assigned a Metabolomics Workbench Project ID: PR000010. The data is directly accessible from The Metabolomics WorkBench database . The data was collected to compare metabolomics, phenotypic and genetic diversity across various groups of Africans. The data consisted of 40 samples; 25 samples from Ethiopia and 15 samples from Tanzania and 5126 metabolites. For the purpose of this study we made sure we had a complete dataset in order to compare the methods. The complete datasets created were two datasets based on the country; Ethiopia dataset (25 samples by 1251 metabolites) and Tanzania dataset (15 samples by 2250 metabolites).

Due to small sample sizes in metabolomics datasets, we used a simulation approach originally designed to resemble the multivariate distribution of gene expression in the original microarray data (Parrish, Spencer Iii et al. 2009). Since our Myocardial Infarction and Atherothrombotic data had missing values we first imputed missing values based on the KNN-CR method and then used the simulation method to simulate 100 datasets. For the African Race data we started with a complete dataset. The different groups were considered as independent datasets and the imputation was done on them separately. We used the similar mechanism for missingness and screening as used in the simulation studies, with sample sizes of 25 and 50 for the myocardial infarction dataset, 50 and 100 for the human atherothrombotic dataset and 15 and 25 for the Tanzania and Ethiopia data sets, respectively, from the African race study.

2.5 Results

Simulation Results

In this section, we present the results of the simulation studies comparing the performance measures of KNN-TN, KNN-CR and KNN-EU. Figures 4, 5 and 6 plot the distribution of the RMSE values for KNN-TN, KNN-CR and KNN-EU by correlation type and percent missing for sample sizes 20, 50, and 100, respectively. Tables 1, 2 and 3 show the average RMSE results of the different simulation settings based on the 100 replications. Since the pattern of results was similar regardless of whether the percent MNAR was less than the percent MAR, results are shown for percent MNAR > percent MAR only. As can be seen from the tables and figures, the results consistently show that the KNN-TN method outperforms both the KNN-CR and KNN-EU methods. ANOVA modeling of the RMSE values shows statistically significant differences between the KNN-TN method and KNN-CR / KNN-EU methods for all three cases, and significant effects for the other two factors (percent missing and correlation type) as well (Tables 4-6). To visualize how our method works we selected a simulated dataset from N = 50 and M = 400 with 15% missing (10% below the LOD and 5% MAR) and compared the true missing values with KNN-TN, KNN-CR and KNN-EU. Figure 7 demonstrates that our imputation method imputes values below the limit of detection whereas the Euclidean or correlation based metrics are less accurate for these values. The figure is reproducible with our included example script in Supplemental File 5. We further compared the three methods with standard imputation methods in metabolomics (zero, minimum and mean imputation methods) and all three KNN imputation algorithms outperformed the standard methods. The results for the simulation studies are shown in Tables 7-9 where we see the average RMSE range was from 4.0 to 5.8.

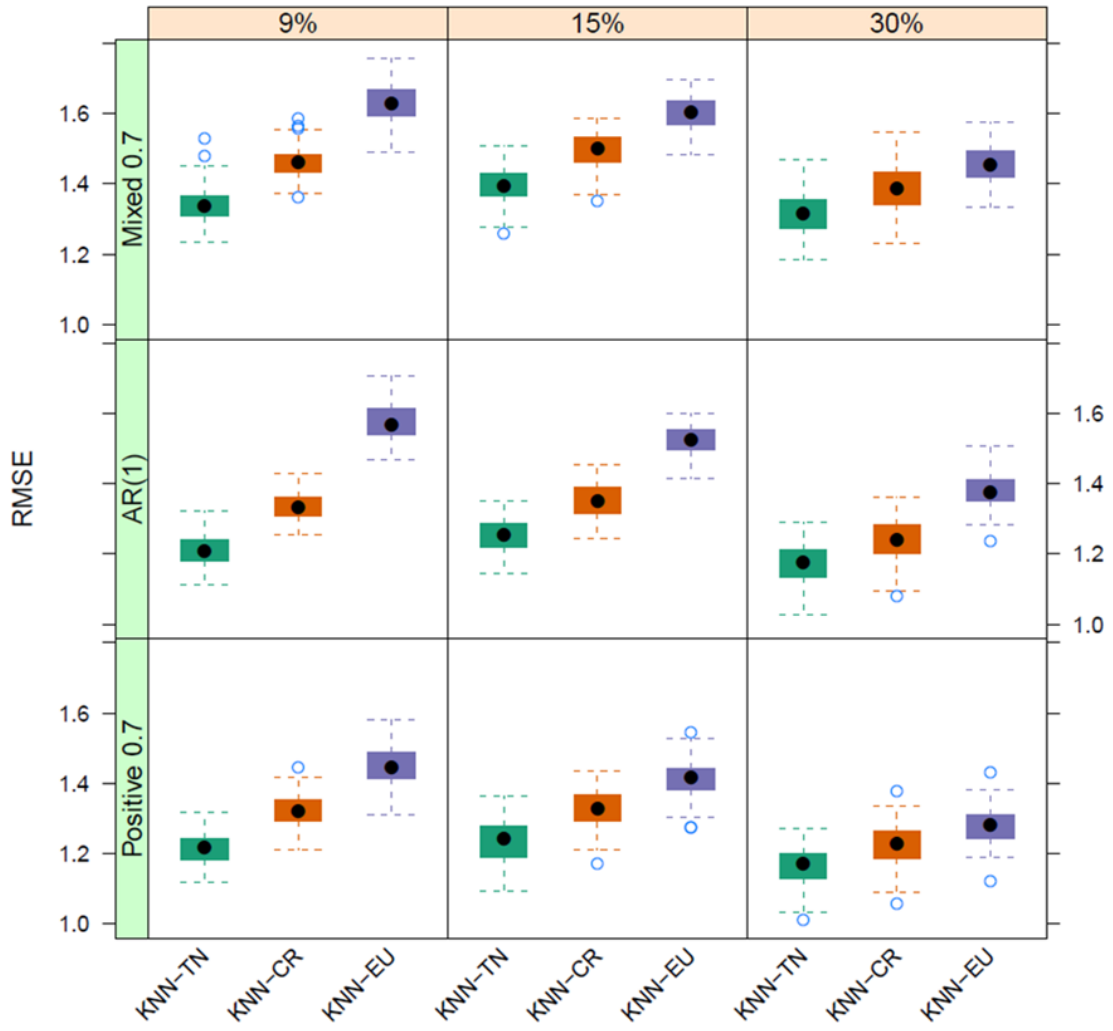

Figure 4: Boxplots of root mean squared error for KNN-TN, KNN-CR and KNN-EU for 100 datasets, 20 samples by 400 metabolites.

Total missing was considered at 9%, 15% and 30% and within each missing MNAR is greater than MAR. The three correlation structure used was i) only positive correlation 0.7, ii) AR(1) correlation 0.9 and iii) mixed correlation 0.7.

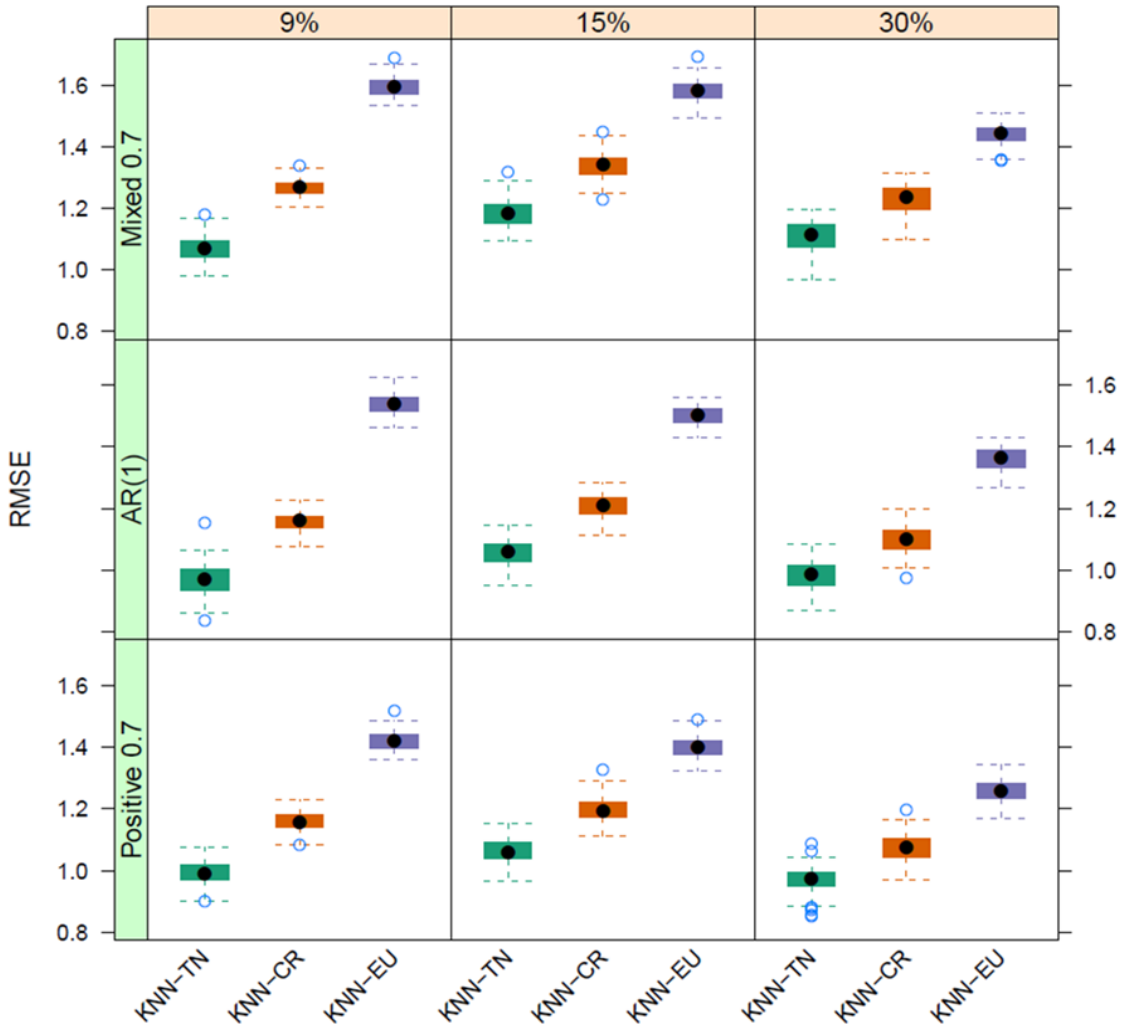

Figure 5: Boxplots of root mean squared error for KNN-TN, KNN-CR and KNN-EU for 100 datasets, 50 samples by 400 metabolites.

Total missing was considered at 9%, 15% and 30% and within each missing MNAR is greater than MAR. The three correlation structure used was i) only positive correlation 0.7, ii) AR(1) correlation 0.9 and iii) mixed correlation 0.7.

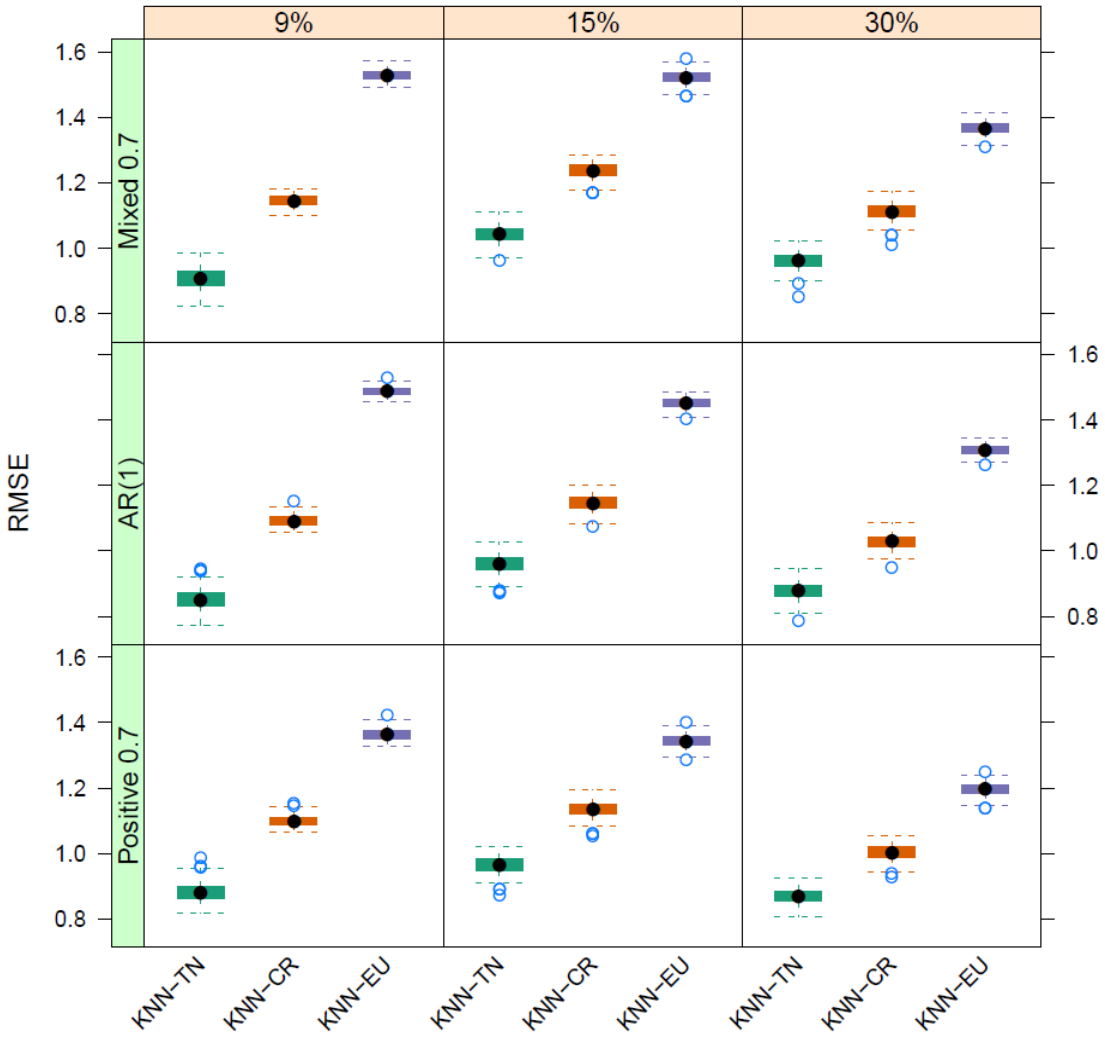

Figure 6: Boxplots of root mean squared error for KNN-TN, KNN-CR and KNN-EU for 100 datasets, 100 samples by 900 metabolites

Total missing was considered at 9%, 15% and 30% and within each missing MNAR is greater than MAR. The three correlation structure used was i) only positive correlation 0.7, ii) AR(1) correlation 0.9 and iii) mixed correlation 0.7.

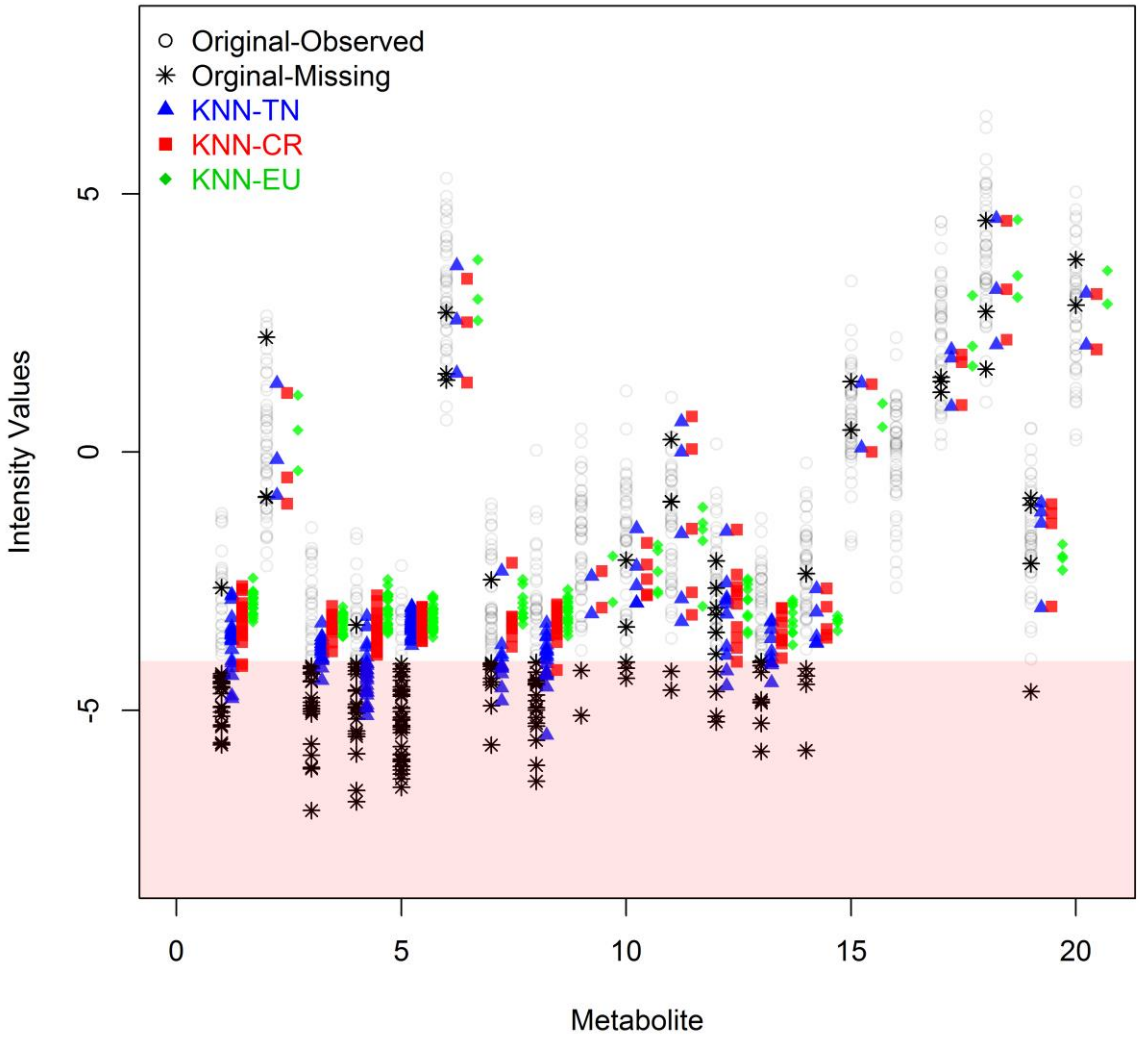

Figure 7: Comparison of the true missing values with missing values imputed from the three methods based on a single simulated dataset (N = 50 X M = 400).

The values for the first 20 metabolites are shown. The x-axis represents the metabolites, and the y-axis represents the intensity values. The open black circles represent observed values, while the black stars represent missing observations. Blue triangles, red squares, and green diamonds represent missing values imputed by KNN-TN, KNN-CR and KNN-EU,

respectively. The region below the LOD is shaded in light red. In most cases, the KNN-TN algorithm is able to impute missing values below the LOD better than the other two methods (e.g., metabolites 1, 3, 4, 7, 8, 12, and 13). In other cases, the KNN-TN imputations are similar to KNN-CR (e.g. for metabolite 5, for which the missing below the LOD was too high and the NR algorithm was unable to converge).

MNAR/MAR	DATA	CORR	KNN-TN	KNN-CR	KNN-EU
6% / 3%	**DATA 1**	**POS 0.7**	1.214 (0.042)	1.321 (0.041)	1.450 (0.051)
	DATA 2	**AR(1) 0.9**	1.210 (0.046)	1.334 (0.039)	1.576 (0.050)
	DATA 3	**MIX 0.7**	1.341 (0.048)	1.462 (0.041)	1.629 (0.053)
10% / 5%	**DATA 1**	**POS 0.7**	1.238 (0.057)	1.325 (0.055)	1.413 (0.053)
	DATA 2	**AR(1) 0.9**	1.251 (0.049)	1.350 (0.048)	1.523 (0.043)
	DATA 3	**MIX 0.7**	1.392 (0.050)	1.495 (0.052)	1.601 (0.050)
20% / 10%	**DATA 1**	**POS 0.7**	1.165 (0.056)	1.226 (0.060)	1.280 (0.051)
	DATA 2	**AR(1) 0.9**	1.172 (0.055)	1.241 (0.057)	1.382 (0.048)
	DATA 3	**MIX 0.7**	1.315 (0.056)	1.385 (0.059)	1.457 (0.054)

Table 1: Average RMSE of 100 datasets, 20 samples by 400 metabolites for KNN-TN, KNN-CR and KNN-EU.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

MNAR/MAR	DATA	CORR	KNN-TN	KNN-CR	KNN-EU
6% / 3%	**DATA 1**	**POS 0.7**	0.992 (0.037)	1.159 (0.028)	1.421 (0.033)
	DATA 2	**AR(1) 0.9**	0.970 (0.050)	1.156 (0.029)	1.539 (0.033)
	DATA 3	**MIX 0.7**	1.071 (0.044)	1.266 (0.027)	1.593 (0.029)
10% / 5%	**DATA 1**	**POS 0.7**	1.062 (0.039)	1.197 (0.038)	1.402 (0.034)
	DATA 2	**AR(1) 0.9**	1.056 (0.045)	1.210 (0.038)	1.502 (0.029)
	DATA 3	**MIX 0.7**	1.184 (0.044)	1.339 (0.041)	1.580 (0.034)
20% / 10%	**DATA 1**	**POS 0.7**	0.969 (0.043)	1.072 (0.043)	1.258 (0.036)
	DATA 2	**AR(1) 0.9**	0.983 (0.047)	1.100 (0.046)	1.361 (0.034)
	DATA 3	**MIX 0.7**	1.110 (0.046)	1.229 (0.045)	1.439 (0.034)

Table 2: Average RMSE of 100 datasets, 50 samples by 400 metabolites for KNN-TN, KNN-CR and KNN-EU.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

MNAR/MAR	DATA	CORR	KNN-TN	KNN-CR	KNN-EU
6% / 3%	**DATA 1**	**POS 0.7**	0.882 (0.030)	1.099 (0.019)	1.364 (0.018)
	DATA 2	**AR(1) 0.9**	0.852 (0.033)	1.092 (0.019)	1.488 (0.014)
	DATA 3	**MIX 0.7**	0.909 (0.031)	1.147 (0.019)	1.530 (0.018)
10% / 5%	**DATA 1**	**POS 0.7**	0.965 (0.029)	1.133 (0.026)	1.344 (0.022)
	DATA 2	**AR(1) 0.9**	0.959 (0.033)	1.145 (0.027)	1.453 (0.018)
	DATA 3	**MIX 0.7**	1.043 (0.027)	1.236 (0.025)	1.521 (0.022)
20% / 10%	**DATA 1**	**POS 0.7**	0.870 (0.024)	1.000 (0.024)	1.197 (0.021)
	DATA 2	**AR(1) 0.9**	0.878 (0.030)	1.028 (0.026)	1.308 (0.016)
	DATA 3	**MIX 0.7**	0.960 (0.027)	1.110 (0.027)	1.368 (0.021)

Table 3: Average RMSE of 100 datasets, 100 samples by 900 metabolites for KNN-TN, KNN-CR and KNN-EU.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

Table 4: Specific differences in RMSE for the imputation methods and ANOVA results for the factors for 20 samples by 400 metabolites.

Contrast	Estimate	Std. Error	t Value	P Value
KNN-CR – KNN-TN	0.094	0.002815	33.303	<2e-16 ***
KNN-EU – KNN-TN	0.224	0.002815	79.519	<2e-16 ***

Table 4 a. Specific differences in RMSE values for the KNN-CR and KNN-EU methods compared to the KNN-TR method for 20 samples by 400 metabolites.

	Df	Sum Sq	Mean Sq	F value	P Value
Imputation Method	2	22.752	11.376	3189.4	<2e-16 ***
Percent Missing	2	6.574	3.287	921.6	<2e-16 ***
Correlation Type	2	12.333	6.166	1728.8	<2e-16 ***
Residuals	2693	9.605	0.004		

Table 4 b. ANOVA table giving the significance of the three factors in the simulation study for 20 samples by 400 metabolites.

Table

Table 5: Specific differences in RMSE for the imputation methods and ANOVA results for the factors for 50 samples by 400 metabolites.

	Estimate	Std. Error	t Value	P Value
KNN-CR – KNN-TN	0.148	0.002517	58.72	<2e-16 ***
KNN-EU – KNN-TN	0.224	0.002517	163.23	<2e-16 ***

Table 5 a. Specific differences in RMSE values for the KNN-CR and KNN-EU methods compared to the KNN-TR method for 50 samples by 400 metabolites.

	Df	Sum Sq	Mean Sq	F value	P Value
Imputation Method	2	77.95	38.98	13672	<2e-16 ***
Percent Missing	2	5.83	2.92	1023	<2e-16 ***
Correlation Type	2	9.71	4.85	1703	<2e-16 ***
Residuals	2693	7.68	0.00		

Table 5 b. ANOVA table giving the significance of the three factors in the simulation study for 50 samples by 400 metabolites.

Table 6: Specific differences in RMSE for the imputation methods and ANOVA results for the factors for 100 samples by 900 metabolites.

	Estimate	Std. Error	t Value	P Value
KNN-CR – KNN-TN	0.186	0.002278	81.60	<2e-16 ***
KNN-EU – KNN-TN	0.473	0.002278	207.60	<2e-16 ***

Table 6 a. Specific differences in RMSE values for the KNN-CR and KNN-EU methods compared to the KNN-TR method for 100 samples by 900 metabolites.

	Df	Sum Sq	Mean Sq	F value	P Value
Imputation Method	2	102.18	51.09	21877	<2e-16 ***
Percent Missing	2	6.55	3.27	1402	<2e-16 ***
Correlation Type	2	5.35	2.67	1145	<2e-16 ***
Residuals	2693	6.29	0.00		

Table 6 b. ANOVA table giving the significance of the three factors in the simulation study for 100 samples by 900 metabolites.

MNAR/MAR	DATA	CORR	Zero	Min	Mean
6% / 3%	**DATA 1**	**POS 0.7**	5.568 (0.164)	5.490 (0.172)	5.539 (0.172)
	DATA 2	**AR(1) 0.9**	5.561 (0.148)	5.482 (0.152)	5.531 (0.152)
	DATA 3	**MIX 0.7**	5.565 (0.159)	5.583 (0.171)	5.626 (0.168)
10% / 5%	**DATA 1**	**POS 0.7**	5.118 (0.143)	5.507 (0.137)	5.118 (0.145)
	DATA 2	**AR(1) 0.9**	5.127 (0.154)	5.066 (0.154)	5.131 (0.154)
	DATA 3	**MIX 0.7**	5.240 (0.145)	5.169 (0.143)	5.229 (0.146)
20% / 10%	**DATA 1**	**POS 0.7**	4.133 (0.149)	4.224 (0.142)	4.251 (0.156)
	DATA 2	**AR(1) 0.9**	4.134 (0.140)	4.229 (0.133)	4.251 (0.145)
	DATA 3	**MIX 0.7**	4.226 (0.132)	4.306 (0.123)	4.335 (0.130)

Table 7: Average RMSE of 100 datasets, 20 samples by 400 metabolites for zero, minimum and mean imputation methods.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

MNAR/MAR	DATA	CORR	Zero	Min	Mean
6% / 3%	**DATA 1**	**POS 0.7**	5.604 (0.131)	5.520 (0.136)	5.569 (0.138)
	DATA 2	**AR(1) 0.9**	5.607 (0.124)	5.526 (0.128)	5.570 (0.127)
	DATA 3	**MIX 0.7**	5.688 (0.115)	5.602 (0.119)	5.651 (0.119)
10% / 5%	**DATA 1**	**POS 0.7**	5.183 (0.136)	5.113 (0.130)	5.176 (0.135)
	DATA 2	**AR(1) 0.9**	5.181 (0.127)	5.109 (0.118)	5.176 (0.121)
	DATA 3	**MIX 0.7**	5.295 (0.127)	5.209 (0.119)	5.279 (0.124)
20% / 10%	**DATA 1**	**POS 0.7**	4.179 (0.144)	4.256 (0.133)	4.294 (0.147)
	DATA 2	**AR(1) 0.9**	4.182 (0.136)	4.256 (0.126)	4.294 (0.137)
	DATA 3	**MIX 0.7**	4.279 (0.135)	4.338 (0.124)	4.387 (0.137)

Table 8: Average RMSE of 100 datasets, 50 samples by 400 metabolites for zero, minimum and mean imputation methods.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

MNAR/MAR	DATA	CORR	Zero	Min	Mean
6% / 3%	**DATA 1**	**POS 0.7**	5.607 (0.083)	5.522 (0.087)	5.570 (0.086)
	DATA 2	**AR(1) 0.9**	5.608 (0.073)	5.524 (0.075)	5.571 (0.076)
	DATA 3	**MIX 0.7**	5.694 (0.085)	5.607 (0.089)	5.654 (0.089)
10% / 5%	**DATA 1**	**POS 0.7**	5.194 (0.104)	5.121 (0.097)	5.186 (0.100)
	DATA 2	**AR(1) 0.9**	5.194 (0.096)	5.110 (0.091)	5.185 (0.092)
	DATA 3	**MIX 0.7**	5.311 (0.091)	5.219 (0.084)	5.291 (0.088)
20% / 10%	**DATA 1**	**POS 0.7**	4.191 (0.091)	4.266 (0.085)	4.305 (0.089)
	DATA 2	**AR(1) 0.9**	4.188 (0.088)	4.266 (0.083)	4.303 (0.087)
	DATA 3	**MIX 0.7**	4.286 (0.092)	4.345 (0.087)	4.394 (0.094)

Table 9: Average RMSE of 100 datasets, 100 samples by 900 metabolites for zero, minimum and mean imputation methods.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

Real Data Simulations Results

We conducted a simulation study based on the real datasets to further validate our results. Table 10, 11, and 12 show the results of the in vivo myocardial infarction data, human atherothrombotic data, and publicly available African Race data. In all cases the KNN-TN and KNN-CR results are substantially better than the KNN-EU results, with RMSE means more than two standard deviations below the means for KNN-EU (p-value < 0.05 for KNN-TN vs. KNN-EU contrast, Tables 13 - 15). The difference between KNN-TN and KNN-CR is much smaller by comparison, with statistically significant differences only for the Atherothrombotic and African Race data sets. However, in every case the mean RMSE for KNN-TN is below that for KNN-CR. Tables 13 – 15 show that significant differences in RMSE values exist according to the other factors in the simulation study (percent missing, group, and sample size) as well. We further compared the three methods the standard imputation methods in metabolomics (zero, minimum and mean imputation methods) and all three KNN imputation algorithms outperformed the standard methods. The results for the real data are shown in Tables 16-18 where we see the average RMSE range was from 2.2 to 7.2. The t-test analysis and the MLCI values are shown in Table 4. A higher value of MLCI indicates that the metabolite list from the imputed data is similar to that from the complete data and from the tables KNN-TN and KNN-CR have the highest values, whereas the KNN-EU, Zero, Minimum and Mean imputation methods have lower MLCI indexes. Differences in mean MLCI values between KNN-TN and KNN-CR were not statistically significant (Tables 19-21), whereas KNN-TN was significantly better than the other four methods in all cases except for the African Race data (where mean imputation and all KNN imputation methods were roughly equivalent and better than zero and minimum value imputation).

MNAR/MAR	SAMPLE SIZE	GROUP	KNN-TN	KNN-CR	KNN-EU
6% / 3%	**25**	**CASES**	0.613 (0.072)	0.619 (0.071)	0.786 (0.075)
	25	**CONTROLS**	0.436 (0.054)	0.441 (0.054)	0.607 (0.047)
	50	**CASES**	0.597 (0.045)	0.602 (0.046)	0.776 (0.048)
	50	**CONTROLS**	0.415 (0.032)	0.420 (0.031)	0.600 (0.028)
10% / 5%	**25**	**CASES**	0.632 (0.099)	0.637 (0.101)	0.810 (0.087)
	25	**CONTROLS**	0.416 (0.052)	0.419 (0.050)	0.555 (0.044)
	50	**CASES**	0.607 (0.073)	0.610 (0.073)	0.809 (0.069)
	50	**CONTROLS**	0.409 (0.034)	0.412 (0.034)	0.556 (0.029)
20% / 10%	**25**	**CASES**	0.610 (0.108)	0.612 (0.107)	0..701 (0.091)
	25	**CONTROLS**	0.381 (0.059)	0.389 (0.058)	0.498 (0.048)
	50	**CASES**	0.586 (0.083)	0.586 (0.081)	0.699 (0.071)
	50	**CONTROLS**	0.370 (0.053)	0.381 (0.053)	0.499 (0.041)

Table 10: Average RMSE of 100 simulations using the in vivo myocardial infarction dataset for KNN-TN, KNN-CR and KNN-EU.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

MNAR/MAR	SAMPLE SIZE	GROUP	KNN-TN	KNN-CR	KNN-EU
6% / 3%	50	sCAD	1.145 (0.047)	1.171 (0.046)	1.410 (0.052)
	50	TYPE1	1.255 (0.054)	1.273 (0.053)	1.555 (0.057)
	50	TYPE2	1.266 (0.051)	1.279 (0.050)	1.567 (0.055)
	100	sCAD	1.083 (0.048)	1.109 (0.041)	1.403 (0.053)
	100	TYPE1	1.183 (0.048)	1.199 (0.041)	1.531 (0.053)
	100	TYPE2	1.183 (0.048)	1.191 (0.041)	1.531 (0.053)
10% / 5%	50	sCAD	1.146 (0.045)	1.168 (0.045)	1.337 (0.050)
	50	TYPE1	1.262 (0.059)	1.280 (0.057)	1.490 (0.059)
	50	TYPE2	1.296 (0.048)	1.315 (0.047)	1.531 (0.051)
	100	sCAD	1.075 (0.031)	1.095 (0.031)	1.330 (0.034)
	100	TYPE1	1.171 (0.039)	1.189 (0.038)	1.460 (0.041)
	100	TYPE2	1.189 (0.040)	1.207 (0.038)	1.490 (0.040)
20% / 10%	50	sCAD	1.120 (0.049)	1.140 (0.049)	1.210 (0.047)
	50	TYPE1	1.261 (0.061)	1.282 (0.061)	1.398 (0.059)
	50	TYPE2	1.354 (0.058)	1.373 (0.058)	1.484 (0.054)
	100	sCAD	1.033 (0.035)	1.053 (0.035)	1.198 (0.034)
	100	TYPE1	1.153 (0.041)	1.176 (0.041)	1.372 (0.041)
	100	TYPE2	1.246 (0.037)	1.266 (0.037)	1.451 (0.036)

Table 11: Average RMSE of 100 simulations using the human atherothrombotic dataset for KNN-TN, KNN-CR and KNN-EU.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

MNAR/MAR	SAMPLE SIZE	DATASET	KNN-TN	KNN-CR	KNN-EU
6% / 3%	**15**	**Tanzania**	0.695 (0.050)	0.711 (0.051)	0.772 (0.049)
	25	**Ethiopia**	0.575 (0.029)	0.592 (0.029)	0.701 (0.033)
10% / 5%	**15**	**Tanzania**	0.659 (0.052)	0.674 (0.053)	0.728 (0.050)
	25	**Ethiopia**	0.556 (0.029)	0.574 (0.029)	0.665 (0.031)
20% / 10%	**15**	**Tanzania**	0.577 (0.049)	0.588 (0.049)	0..627 (0.051)
	25	**Ethiopia**	0.507 (0.026)	0.520 (0.027)	0.599 (0.028)

Table 12: Average RMSE of 100 simulations using the African Race dataset for KNN-TN, KNN-CR and KNN-EU.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

Table 13: Specific differences in RMSE for the imputation methods and ANOVA results for the factors for Myocardial dataset.

	Estimate	Std. Error	t Value	P Value
KNN-CR – KNN-TN	0.005	0.00275	1.737	0.0825
KNN-EU – KNN-TN	0.152	0.00275	55.333	<2e-16 ***

Table 13 a. Specific differences in RMSE values for the KNN-CR and KNN-EU methods compared to the KNN-TR method for Myocardial dataset.

	Df	Sum Sq	Mean Sq	F value	P Value
Imputation Method	2	17.95	8.98	1979.08	<2e-16 ***
Percent Missing	2	1.88	0.94	207.76	<2e-16 ***
Group	1	37.82	37.82	8339.65	<2e-16 ***
Sample Size	1	0.14	0.14	31.56	2.08e-08
Residuals	3593	16.30	0.00		

Table 13 b. ANOVA table giving the significance of the four factors for the Myocardial dataset.

Table 14: Specific differences in RMSE for the imputation methods and ANOVA results for the factors for Atherothrombotic dataset.

	Estimate	Std. Error	t Value	P Value
KNN-CR – KNN-TN	0.019	0.00239	479.375	<2e-16 ***
KNN-EU – KNN-TN	0.240	0.00207	9.261	<2e-16 ***

Table 14 a. Specific differences in RMSE values for the KNN-CR and KNN-EU methods compared to the KNN-TR method for Atherothrombotic dataset.

	Df	Sum Sq	Mean Sq	F value	P Value
Imputation Method	2	64.25	32.13	8330.9	<2e-16 ***
Percent Missing	2	1.65	0.82	213.6	<2e-16 ***
Group	2	27.07	13.54	3510.5	<2e-16 ***
Sample Size	1	5.99	5.99	1554.0	<2e-16 ***
Residuals	5392	20.79	0.00		

Table 14 b. ANOVA table giving the significance of the four factors for the Atherothrombotic dataset.

Table 15: Specific differences in RMSE for the imputation methods and ANOVA results for the factors for African Race dataset.

	Estimate	Std. Error	t Value	P Value
KNN-CR – KNN-TN	0.015	0.002532	5.941	3.4e-09 ***
KNN-EU – KNN-TN	0.087	0.002532	34.523	<2e-16 ***

Table 15 a. Specific differences in RMSE values for the KNN-CR and KNN-EU methods compared to the KNN-TR method for African Race dataset.

	Df	Sum Sq	Mean Sq	F value	P Value
Imputation Method	2	2.621	1.3103	681.4	<2e-16 ***
Percent Missing	2	3.478	1.7388	904.2	<2e-16 ***
Group	1	3.054	3.054	1588.2	<2e-16 ***
Residuals	1794	3.450			

Table 15 b. ANOVA table giving the significance of the four factors for the African Race dataset.

MNAR/MAR	SAMPLE SIZE	GROUP	Zero	Min	Mean
6% / 3%	25	CASES	4.530 (0.175)	3.454 (0.116)	3.474 (0.117)
	25	CONTROLS	4.213 (0.213)	3.234 (0.103)	3.246 (0.097)
	50	CASES	4.556 (0.139)	3.506 (0.090)	3.528 (0.092)
	50	CONTROLS	4.256 (0.158)	3.328 (0.069)	3.339 (0.068)
10% / 5%	25	CASES	4.908 (0.173)	3.262 (0.093)	3.297 (0.094)
	25	CONTROLS	4.704 (0.203)	3.046 (0.082)	3.058 (0.088)
	50	CASES	4.921 (0.142)	3.282 (0.087)	3.315 (0.089)
	50	CONTROLS	4.742 (0.124)	3.098 (0.053)	3.118 (0.053)
20% / 10%	25	CASES	6.053 (0.169)	2.091 (0.076)	2.949 (0.077)
	25	CONTROLS	5.879 (0.158)	2.796 (0.051)	2.803 (0.053)
	50	CASES	6.050 (0.117)	2.916 (0.057)	2.960 (0.056)
	50	CONTROLS	5.900 (0.112)	2.821 (0.037)	2.842 (0.040)

Table 16: Average RMSE of 100 simulations using the in vivo myocardial infarction dataset for zero, minimum and mean imputation methods.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

MNAR/MAR	SAMPLE SIZE	GROUP	Zero	Min	Mean
6% / 3%	50	sCAD	5.437 (0.084)	4.331 (0.066)	4.363 (0.067)
	50	TYPE1	5.608 (0.087)	4.543 (0.080)	4.580 (0.082)
	50	TYPE2	5.652 (0.093)	4.547 (0.094)	4.588 (0.093)
	100	sCAD	5.429 (0.053)	4.329 (0.045)	4.362 (0.046)
	100	TYPE1	5.569 (0.057)	4.463 (0.058)	4.500 (0.059)
	100	TYPE2	5.629 (0.062)	4.504 (0.061)	5.542 (0.062)
10% / 5%	50	sCAD	5.876 (0.081)	4.066 (0.051)	4.117 (0.052)
	50	TYPE1	5.988 (0.083)	4.188 (0.073)	4.245 (0.077)
	50	TYPE2	6.056 (0.070)	4.216 (0.066)	4.273 (0.068)
	100	sCAD	5.891 (0.052)	4.079 (0.035)	4.131 (0.036)
	100	TYPE1	6.013 (0.055)	4.216 (0.056)	4.270 (0.059)
	100	TYPE2	6.066 (0.051)	4.216 (0.056)	4.273 (0.057)
20% / 10%	50	sCAD	7.001 (0.064)	3.519 (0.041)	3.581 (0.042)
	50	TYPE1	7.141 (0.070)	3.613 (0.052)	3.686 (0.056)
	50	TYPE2	7.206 (0.070)	3.716 (0.052)	3.801 (0.057)
	100	sCAD	7.019 (0.048)	3.550 (0.032)	3.612 (0.034)
	100	TYPE1	7.152 (0.050)	3.638 (0.046)	3.711 (0.050)
	100	TYPE2	7.207 (0.043)	3.705 (0.041)	3.786 (0.046)

Table 17: Average RMSE of 100 simulations using the human atherothrombotic dataset for zero, minimum and mean imputation methods.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

MNAR/MAR	SAMPLE SIZE	GROUP	Zero	Min	Mean
6% / 3%	**15**	**Tanzania**	4.114 (0.167)	2.752 (0.078)	2.764 (0.078)
	25	**Ethiopia**	4.022 (0.131)	2.673 (0.055)	2.685 (0.053)
10% / 5%	**15**	**Tanzania**	4.567 (0.173)	2.570 (0.074)	2.579 (0.075)
	25	**Ethiopia**	4.535 (0.114)	2.489 (0.048)	2.499 (0.050)
20% / 10%	**15**	**Tanzania**	5.862 (0.119)	2.293 (0.052)	2.288 (0.057)
	25	**Ethiopia**	5.813 (0.108)	2.255 (0.035)	2.249 (0.037)

Table 18: Average RMSE of 100 simulations using the African Race dataset for zero, minimum and mean imputation methods.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR

Table 19: Specific differences in MLCI for the imputation methods and ANOVA results for the factors for Myocardial Infarction dataset.

Contrast	Estimate	Std. Error	t Value	P Value
Zero – KNN-TN	-0.511	0.0046	-96.82	<2e-16 ***
Min – KNN-TN	-0.152	0.00527	-28.75	<2e-16 ***
Mean – KNN-TN	-0.038	0.00527	-7.23	5.8e-13
KNN-CR – KNN-TN	-0.002	0.00527	-0.41	0.679
KNN-EU – KNN-TN	-0.024	0.00527	-4.42	1.03e-05

Table 19 a. Specific differences in MLCI values for the Zero, minimum, mean, KNN-CR and KNN-EU methods compared to the KNN-TR method for the Myocardial Infarction data.

	Df	Sum Sq	Mean Sq	F value	P Value
Imputation Method	5	118.71	23.742	2846.5	<2e-16 ***
Percent Missing	2	8.81	4.407	528.4	<2e-16 ***
Sample Size	1	3.18	3.178	381.0	<2e-16 ***
Residuals	3591	29.95	0.008		

Table 19 b. ANOVA table giving the significance of the three factors in the simulation study for Myocardial Infarction data.

Table 20: Specific differences in MLCI for the imputation methods and ANOVA results for the factors for Atherothrombotic dataset.

Contrast	Estimate	Std. Error	t Value	P Value
Zero – KNN-TN	-0.295	0.0044	.-58.47	<2e-16 ***
Min – KNN-TN	-0.112	0.0050	-22.12	<2e-16 ***
Mean – KNN-TN	-0.036	0.0050	-7.17	8.9e-13
KNN-CR – KNN-TN	-0.001	0.0050	0.25	0.803
KNN-EU – KNN-TN	-0.017	0.0050	-3.36	0.0008

Table 20 a. Specific differences in MLCI values for the Zero, minimum, mean, KNN-CR and KNN-EU methods compared to the KNN-TR method for the Atherothrombotic data.

	Df	Sum Sq	Mean Sq	F value	P Value
Imputation Method	5	39.67	7.93	1037.9	<2e-16 ***
Percent Missing	2	2.08	1.04	136.1	<2e-16 ***
Sample Size	1	3.10	3.10	405.0	<2e-16 ***
Residuals	3591	27.45	0.01		

Table 20 b. ANOVA table giving the significance of the three factors in the simulation study for Atherothrombotic data.

Table 21: Specific differences in MLCI for the imputation methods and ANOVA results for the factors for African Race dataset.

Contrast	Estimate	Std. Error	t Value	P Value
Zero – KNN-TN	-0.239	0.0080	-24.523	<2e-16 ***
Min – KNN-TN	-0.139	0.0097	-14.214	<2e-16 ***
Mean – KNN-TN	-0.008	0.0097	-0.849	0.396
KNN-CR – KNN-TN	-0.001	0.0097	0.052	0.958
KNN-EU – KNN-TN	-0.003	0.0097	-0.262	0.795

Table 21 a. Specific differences in MLCI values for the Zero, minimum, mean, KNN-CR and KNN-EU methods compared to the KNN-TR method for the African Race data.

	Df	Sum Sq	Mean Sq	F value	P Value
Imputation Method	5	15.39	3.079	216.1	<2e-16 ***
Percent Missing	2	1.06	0.530	37.2	<2e-16 ***
Residuals	1792	25.53	0.014		

Table 21 b. ANOVA table giving the significance of the three factors in the simulation study for Myocardial Infarction data.

2.6 Discussions

The objective of this study was to develop an approach for imputing missing values in data generated by mass spectrometry. When metabolites occur at low abundance, below the detection limit of the instrumentation, we can consider it as missing not at random. In contrast, missing values resulting from technical errors are considered missing at random. To this end, we introduce an extension to the KNN imputation algorithm which handles truncated data, termed KNN-TN. To our knowledge, this is the first proposal approach which can simultaneously handle missing data generated by both MNAR (falling below the LOD) and MAR mechanisms. Since MNAR is involved and is due to the detection limit, we consider the detection limit as a truncation point and assume that the metabolite follows a truncated normal distribution. Therefore the mean and standard deviation are estimated from the truncated normal distribution and used to standardize the metabolites in the KNN imputation algorithm. The simulation results show that the proposed method performs better than KNN based on correlation or Euclidean measures when there is missing data due to a threshold LOD.

In our simulations we evaluated three different data set sizes: small (20 samples by 400 metabolites), medium (50 samples by 400 metabolites) and large (100 samples by 900 metabolites). As the sample size increased, the RMSE was lower for the different missing percentages. The LOD was calculated based on the missing percentage. For instance in 9% missing (where 6% was considered as MNAR) the 6% quantile for the complete data was considered as the LOD where we considered everything below that value as missing. For the simulation studies, the results shown in the tables are based on when the MNAR percentage is greater than the MAR percentage (e.g. for 9% total missing, 6% is MNAR and 3% is MAR). However the results were similar when the MAR percentage was greater than the MNAR

percentage, with KNN-TN outperforming both KNN-CR and KNN-EU. In our results, when MNAR is greater than MAR we typically observed the RMSE was greatest at 15% MVs whereas it was lowest at 30% MVs. This counter-intuitive result is likely due to the fact that in the cleaning process (which removes metabolites with >75% MVs) we are removing more metabolites whose values are concentrated near the LOD. For example in the case of 50 samples by 400 metabolites, after screening we reduced the metabolites to an average of about 387 metabolites for 15% missing and 345 metabolites for 30% missing. When the MAR was greater than MNAR, the RMSE increased with the increase in MV percentage.

Troyanskaya et al. (Troyanskaya, Cantor et al. 2001) evaluated a number of different missing value imputation methods and suggested the KNN method to be more robust and sensitive compared to the other methods. In another study by Brock et al (Brock, Shaffer et al. 2008), they compared the KNN based on two different neighbor selection metrics, Euclidean and Correlation and concluded that the correlation based neighbor selection performed better than the Euclidean neighbor selection in most of the cases. In this study we focused on enhancing the KNN method specifically for imputing values when there is missing due to an LOD. Future studies will evaluate how these methods compare to other imputation algorithms in this setting.

Recently, several studies have investigated imputation for MS data (Hrydziuszko and Viant 2011, Gromski, Xu et al. 2014, Taylor, Ruhaak et al. 2016). Taylor et al. (Taylor, Ruhaak et al. 2016) evaluated seven different imputation methods (half minimum, mean, KNN, local least squares regression, Bayesian principal components analysis, singular value decomposition and random forest) and its effects on multiple biological matrix analyses, more specifically on the within-subject correlation of compounds between biological matrices and its consequences on

MANOVA results. They concluded that no imputation method was superior but the mean and half minimum performed poorly. Gromski et al (Gromski, Xu et al. 2014) looked at five different imputation methods (zero, mean, median, KNN and random forest) and its influence on unsupervised and supervised learning. Their results recommended that random forest is better than the other imputation methods and it provided better results in terms of classification rates for both principal components-linear discriminant analysis and partial least squares-discriminant analysis. Hrydziuszko et al. (Hrydziuszko and Viant 2011) suggested the need of missing value imputation as an important step in the processing pipeline. They used metabolomics datasets based on infusion Fourier transform ion cyclotron resonance mass spectrometry and compared eight different imputation methods (predefined value, half minimum, mean, median, KNN, Bayesian Principal Component Analysis, Multivariate Imputation, and REP). Based on their findings, KNN performed better than the other methods.

We included a preliminary investigation of the impact of MV imputation on downstream statistical analysis of metabolomics data. While the KNN-TN method was significantly better than four other imputation algorithms (zero imputation, minimum value imputation, and KNN-EU imputation) in two of three data sets, it was no better than KNN-EU imputation. Further, on the African Race data set there was no significant difference between any of the KNN imputation algorithms and mean imputation, though all were better than zero and minimum value imputation. Although this result is somewhat disappointing, a more comprehensive study of all potential downstream analyses is needed to fully determine, whether the improved imputation accuracy of the KNN-TN method translates into better downstream statistical analysis, and the characteristics of data sets for which more advanced imputation algorithms offer a decided advantage (Oh, Kang et al. 2010).

In some cases (high percent missing or small sample size) the variability of the RMSE for KNN-TN is higher than or similar to that for KNN-CR. This is directly related to the estimation of the mean and variance for the truncated normal distribution, which can be difficult when there are excessive amounts of missing data. In fact, for sample sizes less than 20 there is little to no gain in using KNN-TN over KNN-CR, unless the missing percentage is below the values evaluated in this study (data not shown). To stabilize the estimation of these parameters, one possibility is to again borrow information from metabolites having similar intensity profiles. This is akin to the empirical Bayes approach used to fit linear models and generalized linear models in microarray and RNA-seq studies (Smyth , Smyth 2005, Robinson, McCarthy et al. 2010, Anders and Huber 2012). Our future research will explore this possibility for improving the KNN-TN algorithm.

A related limitation is the reliance on the normality assumption for estimating the truncated mean and standard deviation. In our simulation study we investigated data from a normal distribution, whereas in many cases metabolite data will be non-normally distributed. In these cases we suggest to first transform the data to normality, then impute the values and lastly transform back. As seen in our real datasets, the metabolites are not normally distributed and we log transform them to approximately achieve normality prior to imputation.

The likelihood used in our KNN-TN method is based solely on the observed metabolite data. The full data likelihood would include missing data as well. This is difficult to specify in the current situation as the mechanism by which the MVs were generated (e.g., MNAR, MAR, or MCAR) is unknown. It is possible to improve the algorithm by incorporating these MVs directly into the likelihood function, but ancillary information (e.g., from metabolites

determined to be neighbors) is necessary to inform the system regarding the missingness mechanism (e.g., via the EM-algorithm).

2.7 Conclusions

In conclusion, the experimental results reveal that compared with KNN based on correlation and Euclidean metrics, KNN based on truncation estimation is a competitive approach for imputing high dimensional data where there is potential missingness due to a truncation (detection) threshold. Results based on both real and simulated experimental data show that the proposed method (KNN-TN) generally has lower RMSE values compared to the other two KNN methods and simpler imputation algorithms (zero, mean, and minimum value imputation) when there is both missing at random and missing due to a threshold value. Assessment based on concordance in statistical significance testing demonstrate that KNN-TN and KNN-CR are roughly equivalent and generally outperform the other four methods. However, the approach has limitations with smaller sample sizes, unless the missing percentage is also small. Lastly, even though this study is based on metabolomic datasets our findings are more generally applicable to high-dimensional data that contains missing values associated with an LOD, for instance proteomics data and delta-CT values from qRT-PCR array cards (Warner, Mukhopadhyay et al. 2014).

CHAPTER 3
BAYESIAN APPROACH FOR IMPUTATION OF MISSING VALUES WITH APPLICATION TO HIGH DIMENSIONAL DATA WITH DETECTION LIMIT THRESHOLD

3.1 Background

In many typical high throughput studies, a large number of features (genes/proteins/transcriptomes/metabolites) are measured quantitatively from biological samples, e.g. from humans or animals. Metabolomics is the most downstream field in the omics cascade and provides vital information about metabolic pathways and significant biomarkers related to a certain phenotype. Since metabolites are downstream products, they are very sensitive to various biological states, and can potentially be more readily used for early disease detection than other molecular information, as well as provide contemporaneous information for a variety of other studies (Xi et al 2015). In most MS studies, the number of features (p) is much larger than the number of samples (n). Because of this large p and small n, one of the obstacles is to avoid over-fitting the data. Bayesian methods have become widespread in numerous scientific fields, and this rapid rise in popularity is partially attributable to the decrease in the cost of computational assets that are needed to estimate more complex models (Daniel 2016, Dunson, 2001). There have been several Monte Carlo simulation studies and recent methodologies that have illustrated the benefits of Bayesian methods over frequentist maximum likelihood (ML) methods with small sample sizes (Daniel 2016, Depaoli

& van de Schoot, 2015, Depaoli and Clifton Dunson, 2000, McNeish & Stapleton, 2016). Bayesian statistics is used mainly when complex models cannot be estimated using conventional statistics (Depaoli and Schoot 2016), and many complex models require Bayesian methods to improve convergence problems (Depaoli & Clifton, 2015). It is not necessarily based on large samples and can produce practical results with moderate to small samples, especially when strong prior information is available. Also with the prior distribution, one can utilize the un(certainty) about a parameter and update this knowledge (Depaoli and Schoot 2016).

Many studies have noted the use of Bayesian methods over frequentist methods to better accommodate small sample sizes due to the promise of not using sample size adjustments (Doron & Gaudreau, 2014, Kliem, Kröger, & Kosfelder, 2010, Stenling, Ivarsson, Johnson, & Lindwall, 2015). Consistency and asymptotic normality are necessary properties for inference based on the Maximum Likelihood (ML) estimation, a common frequentist approach. These properties require large sample sizes, and thus, ML estimates in smaller sample sizes can be quite poor (Lee & Song, 2004; McNeish & Stapleton, 2014).

In this work, we develop a Bayesian model for imputing missing values with small sample sizes. The model is based on data augmentation, a common estimation technique in missing value problems (Tanner and Wong 1987). It has been widely used as an alternative to the EM algorithm (Dempster et al 1977) and maximum likelihood estimation. The most commonly used approach for fitting hierarchical models to data is based on the Bayesian paradigm (Lele et al, 2007, Link et al 2002, Clark 2005, Clark and Gelfand 2006). Computing the Bayesian posterior distribution for models became feasible with the advent of the Markov chain Monte

Carlo (MCMC) algorithms (Lele et al 2007). We develop an MCMC algorithm to estimate the posterior distribution of the parameters in our models.

3.2 Methods

We develop a Bayesian model and MCMC algorithm for estimating the posterior distribution of the parameters in our model via a Gibbs sampler. The idea in Gibbs sampling is to generate posterior samples by sweeping through each variable (or block of variables) to sample from its conditional distribution with the remaining variables fixed to their current values. The algorithm below shows the general framework for the Gibbs sampler.

Gibbs sampler algorithm

Intialize $x^{(0)} \sim q(x)$

for iteration $i = 1,2, \dots, M$ **DO**

$$x_1^{(i)} \sim p\left(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, \dots, X_N = x_N^{(i-1)}\right)$$

$$x_2^{(i)} \sim p\left(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, \dots, X_N = x_N^{(i-1)}\right)$$

$$\dots$$

$$x_N^{(i)} \sim p\left(X_N = x_N | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, \dots, X_{N-1} = x_{N-1}^{(i)}\right)$$

End iteration

This process continues until convergence is attained, and the sampling is not done directly from the full posterior distribution but rather sweeping through all the posterior conditionals, one variable block at a time. MCMC algorithms are typically run for a large number of iterations assuming convergence to the target posterior is achieved. The theory of MCMC guarantees that the stationary distribution of the samples generated is the target joint posterior that we are interested in. Due to the impact of initial values, the samples simulated based on this algorithm at early iterations may not necessarily be representative of the actual posterior

distribution. It is common to discard these early samples, and the discarded iterations are often referred to as the "burn-in" period. However, there are several limitations to Gibbs sampling. First, even if we have the full posterior joint density function, it may not be possible to derive the conditionals for each variable in the model. Secondly, if we have the posterior conditionals for each variable, it might be that they are not a known form, and therefore, there is not a direct method to generate samples. The Metropolis-Hastings (MH) algorithm simulates samples from a probability distribution by making use of the full joint density function and a proposal distribution for each variable of interest. Unlike Gibbs samples, the MH algorithm doesn't require the ability of generating samples from all the full conditional distributions, but a proposal or candidate distribution is chosen given the current value of the random variables. The algorithms below shows the general framework for the MH algorithm.

Metropolis-Hastings algorithm

Initialize $x^{(0)} \sim q(x)$

for iteration $i = 1,2, \dots, M$ **DO**

Propose: $x^{cand} \sim q\left(x^{i} | \, x^{(i-1)}\right)$

Acceptance Probability:

$$\alpha\left(x^{cand} | \, x^{(i-1)}\right) = \min\left\{1, \frac{q\left(x^{(i-1)} | x^{cand}\right) \pi\left(x^{cand}\right)}{q\left(x^{cand} | x^{(i-1)}\right) \pi\left(x^{(i-1)}\right)}\right\}$$

$u \sim \text{Uniform}(u;\ 0,1)$

if $u < \alpha$ **then**

Accept the proposal: $x^{(i)} = x^{cand}$

else

Reject the proposal: $x^{(i)} = x^{(i-1)}$

end if

End Iteration

The Bayesian paradigm historically has been proven to stabilize parameter estimation by borrowing information available from similar features (/ proteins/metabolites). Let $Y_i \sim N(X_i\beta, \Sigma)$, where Y_{ij} is the intensity of metabolite $j\ (1 \leq j \leq M)$ in the sample $i\ (1 \leq i \leq N)$ and X_i is the design matrix for the sample. The vector of the mean parameters β will contain the mean value of each metabolite, as well as group/treatment differences as determined by the design matrix. As M is much larger than N, we adopt a lower-dimensional structure on the covariance matrix Σ with autoregressive correlation structure. The Σ is a function of $p + 1$ parameters having a structured form $\Sigma = \Sigma\,(\sigma_1^2, \ldots, \sigma_M^2, \rho)$. We develop a robust approach which allows for both MVs due to MNAR (missing below the LOD threshold ξ) and MAR (missing for a different reason). The MNAR framework requires specification of the missing data mechanism (MDM). Letting R_{ij} be equal to 0 if Y_{ij} is missing and 1 if Y_{ij} is observed, our MDM has the form

$$\Pr\left(R_{ij} = 0 \,\middle|\, Y_{ij}\right) = \begin{cases} \alpha, & Y_{ij} > \xi \\ 1, & Y_{ij} \leq \xi \end{cases}.$$

If the value of Y_{ij} is less than the threshold, it will always be seen to as missing (due to MNAR), and if $Y_{ij} > \xi$, then Y_{ij} may be missing with probability α (due to MAR). We use non-informative priors. A conjugate normal prior with large variance for the regression coefficients β is used, along with a hierarchical structure for the metabolite-specific variances to allow sharing across metabolites and stabilization of estimates. The priors for the parameters are as follows:

$$\alpha \sim \text{Uniform}\ (0,1)$$

$$\beta \sim \text{MVN}(0, 100^2)$$

$$\log(\sigma_j^2) \sim \text{Normal } (\lambda_1, \lambda_2)$$

$$\lambda_1 \sim \text{Normal } (a, b)$$

$$\lambda_2 \sim \text{Inv Gamma } (c, d)$$

$$\rho \sim \text{Uniform } (0,1)$$

Combining the data likelihood, MDM and prior distributions, the full joint density is proportional to

$$P(R_{ij}, Y_{ij}, \beta, \alpha, \sigma_j^2) \propto \prod_{ij}\left\{(1-\alpha)^{R_{ij}\, I(Y_{ij}>\xi)} * \alpha^{(1-R_{ij})I(Y_{ij}>\xi)}\right\} \times$$

$$\prod_i \left\{|\Sigma|^{-\frac{1}{2}} * exp\left\{-\frac{1}{2}(Y_i - X_i\beta)'|\Sigma|^{-1}(Y_i - X_i\beta)\right\}\right\} \times$$

$$exp\left\{-\frac{1}{2}\beta'^{\Omega^{-1}}\beta\right\} \times I(0 < \alpha < 1) \times \prod_j \left\{(\sigma^2)^{-\lambda_1 - 1} * \exp\left\{\frac{-\lambda_2}{\sigma_j^2}\right\}\right\} \times$$

$$I(0 < \rho < 1) \times \frac{1}{\sqrt{2b^2}} \exp\left\{-\frac{1}{2b^2}(\lambda_1 - a)\right\} \times \frac{d^c}{\Gamma(c)} \lambda_2^{c-1} \exp\left\{-\frac{d}{\lambda_2}\right\}$$

Based on MCMC Gibbs sampling from the above joint density, we can obtain full conditional distributions for each parameter.

The conditional densities are:

$$P(\alpha \mid \ldots) \sim \text{Beta}\left(\Sigma_{ij}\left[(1-R_{ij})I(Y_{ij} > \xi)\right] + 1, \Sigma_{ij}\left[R_{ij}I(Y_{ij} > \xi)\right] + 1\right)$$

$$P(\beta \mid \ldots) \sim \text{MVN } ((\Sigma_i X_i'\Sigma^{-1}X_i + \Omega^{-1})^{-1}(\Sigma_i X_i'\Sigma^{-1}Y_i), (\Sigma_i X_i'\Sigma^{-1}X_i + \Omega^{-1})^{-1})$$

$$P(\lambda_1 \mid \ldots) \sim \text{Normal } \left(\frac{\lambda_2 a + b\sum \log \sigma_j^2}{\lambda_2 + bp}, \frac{b\lambda_2}{\lambda_2 + bp}\right)$$

$$P(\lambda_2 \mid \ldots) \sim \text{Inv Gamma } \left(c + \frac{M}{2}, \frac{d + \sum (\log \sigma_j^2 - \lambda_1)^2}{2}\right)$$

The conditional densities for σ_j^2 and ρ are not in the closed form but are proportional to

$$P\left(\sigma_j{}^2 \mid \ldots\right) \propto \prod(\sigma_j^2 \mid \lambda_1, \lambda_2) \times \prod_{i=1}^{N} f(Y_i \mid X_i\beta, \Sigma(\sigma_j^2, \ldots))$$

$$P(\rho \mid \ldots) \propto \prod_{i=1}^{N} f(Y_i \mid X_i\beta, \Sigma(\rho, \ldots) \times \mathrm{I}(0 < \rho < 1)$$

These parameters are sampled using MH steps with the proposal distributions

$$\sigma_j^2 \sim \log \text{Normal } (\log(\sigma_j^2)^{(i-1)}, h)$$

$$\rho \sim \text{Trunc Normal } \left(\rho^{(i-1)}, k\right) \mathrm{I}\ (0 < \rho < 1)$$

where $(\sigma_j^2)^{(i-1)}$ and $\rho^{(i-1)}$ represent the values of the parameters at the previous iteration$(i-1)$. The values of h and k are determined by trial and error so that the proposed value is accepted 25%-45% of the time.

The MCMC sampler depends on data augmentation designed for our MDM to handle the MVs in our data. In each iteration of the Markov chain, the missing Y_{ij}s are imputed based on values of all current model parameters. The proposed MDM assures simple and efficient conjugate sampling for this step. For each missing Y_{ij}, an indicator variable Z_{ij} is introduced which determines whether Y_{ij} will be below the LOD threshold ξ $(Z_{ij} = 1)$ or above the threshold $(Z_{ij} = 0)$. Conditional on the parameter values and the values of $Y_{i(j')}$ for the other metabolites j', Z_{ij} is sampled according to

$$\Pr\left(Z_{ij} = 1\right)$$

$$= \frac{\int_{-\infty}^{\xi} \left(2\pi\tilde{\sigma}_j^2\right)^{-1/2} exp\left\{\frac{-1}{2\tilde{\sigma}_j^2}\left(y - \tilde{\mu}_{ij}\right)^2\right\} dy}{\int_{-\infty}^{\xi} \left(2\pi\tilde{\sigma}_j^2\right)^{-1/2} exp\left\{\frac{-1}{2\tilde{\sigma}_j^2}\left(y - \tilde{\mu}_{ij}\right)^2\right\} dy + \alpha \int_{\xi}^{\infty} \left(2\pi\tilde{\sigma}_j^2\right)^{-1/2} exp\left\{\frac{-1}{2\tilde{\sigma}_j^2}\left(y - \tilde{\mu}_{ij}\right)^2\right\} dy},$$

where $\tilde{\mu}_{ij}$ and $\tilde{\sigma}_j^2$ represent the usual formulas for the conditional mean and variance for Y_{ij} given the $Y_{i(j')}$s in the multivariate normal distribution.

$$\tilde{\mu}_{ij} = \mu_{ij} + \Sigma_{12}\Sigma_{22}^{-1}(Y_{i(j')} - \mu_{i(j')}) \text{ and } \tilde{\sigma}_j^2 = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

As the integrals required are all probabilities under the normal distribution, this can be efficiently evaluated. Given Z_{ij}, the value of the missing Y_{ij} is imputed by

- $Z = 1$: $Y_{ij} \sim$ Truncated Normal $\left(\tilde{\mu}_{ij}, \tilde{\sigma}_j^2\right) I(-\infty, \xi)$
- $Z = 0$: $Y_{ij} \sim$ Truncated Normal $\left(\tilde{\mu}_{ij}, \tilde{\sigma}_j^2\right) I(\xi, \infty)$

Our MCMC scheme iterates between the above steps for a 2300 iterations after a burn-in of 200 iterations. Inference is based on either the posterior sample of model parameters. Alternatively, one or more completed data sets can be formed using the sampled MVs which can be analyzed using existing (frequentist) methodology.

3.3 Simulation Studies

We evaluated the performance of the methods by first evaluating the estimation error. We compared the bias and the MSE for the β regression coefficients based on the Bayesian method and other standard approaches such as zero, mean and minimum imputation. For the standard approaches a general linear regression was fit to estimate the β. We further explored the null hypothesis significance testing, where the primary goal is to determine whether a particular "null" value of a parameter can be rejected. We compare the Bayesian approach to a general t-test under the alternative imputation choices. The above algorithm produces full Bayesian posterior samples, and we need to perform inference on the intercept and treatment effects. Typically, the main interest is differential expressed metabolites between the groups.

The design matrix is chosen such that this is represented as the decision between $\beta_j = 0$ or $\beta_j \neq 0$ for some *j*. We then form $100(1-\alpha)\%$ equal tail credible intervals and check if it contains zero or not. We can obtain a "Bayesian p-value" by taking the largest α such that the $100(1-\alpha)\%$ CI contains zero, that is, the narrowest CI that contains zero. We then evaluate the power, type1 error and the area under the ROC curve (AUC) using the p-values from each test to analyze differentially expressed metabolites.

The simulations were conducted with 100 replicated datasets and are similar in spirit to those used in Tutz and Ramzan (2015) . For each replication we generated data with sample sizes n =10 belonging to 2 groups of 5 samples and number of metabolites m = 225. The first 100 metabolites were differentially expressed whereas the remaining 125 were not differentially significant. The correlation matrix of metabolites within samples was considered to be autoregressive (AR) type correlation, where an AR correlation matrix of order one is defined by pairwise correlations $\rho^{|i-j|}$, for metabolites $i, j = 1, \ldots, M$.

The means of group 1 metabolites generated from a $\text{Uniform}(-5,5)$ distribution. The differentially expressed metabolites (100 out of 225) are one unit larger in group 2. In the AR type correlation, the correlation ρ = 0.8 and the variance was 1. For the degree of missing, three levels were studied: 9% missing, 15% missing and 30% missing. Missing data were created based on the two kinds of missingness, MNAR and MAR. Technically, the latter are generated by MCAR, though a MAR mechanism can be exploited for imputation since the metabolite values are highly correlated. Within each level of missing, a one-third and two-third combination was used to create both MNAR and MAR. For example in 9% missing, we considered 6% as MNAR and 3% as MAR. Data below the given MNAR percentile was considered as missing and the MAR percent was randomly generated in the non-missing data.

The datasets with missing values were passed through a cleaning process where metabolites with more than 50% missing observations were eliminated individually.

3.4 Results

In this section, we present the results of the simulation studies comparing the performance of the Bayesian method with the three standard approaches. Table 22 shows the average bias and MSE based on 100 replications. We consider the accuracy for the intercept, treatment effect across the differentially expressed metabolites (real value =1) and the treatment effect across the non-differentially expressed metabolites (real value =0). As can be seen, the results show that the MSE for the Bayesian method is lower compared to the other methods for the intercepts and the treatment effects. Table 23 shows the average power, type1 error and the AUC based on the Bayesian method and the standard approaches. The power and AUC for the Bayesian method is higher compared to the other methods whereas the Type 1 error is not lower than the compared methods. Figures 8 and 9 plots the distribution of the Bias and MSE values for Bayes, zero, mean and minimum by percent missing for the intercepts, significant treatment effect and non-significant treatment effect. Figure 10 plots the results from the hypothesis testing as the distribution of power, type 1 error and AUC by percent missing.

MNAR (MAR)	β	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
		Bayes		Zero		Min		Mean	
6% (3%)	Int	0.008 (0.095)	0.203 (0.039)	0.115 (0.093)	0.438 (0.065)	-0.137 (0.099)	0.404 (0.068)	0.145 (0.097)	0.484 (0.076)
	S.TE	-0.026 (0.192)	0.406 (0.123)	-0.160 (0.170)	0.610 (0.129)	-0.063 (0.195)	0.838 (0.191)	-0.156 (0.171)	0.622 (0.133)
	NS.TE	0.017 (0.137)	0.386 (0.098)	0.014 (0.126)	0.480 (0.094)	0.005 (0.145)	0.714 (0.144)	0.038 (0.129)	0.506 (0.097)
10% (5%)	Int	0.014 (0.096)	0.206 (0.038)	0.074 (0.098)	0.442 (0.065)	-0.240 (0.103)	0.554 (0.088)	0.141 (0.102)	0.503 (0.082)
	S.TE	-0.033 (0.192)	0.409 (0.122)	-0.176 (0.175)	0.644 (0.162)	-0.094 (0.199)	1.06 (0.241)	-0.168 (0.178)	0.636 (0.161)
	NS.TE	0.015 (0.134)	0.388 (0.102)	0.022 (0.130)	0.494 (0.097)	0.032 (0.162)	0.907 (0.180)	0.051 (0.130)	0.505 (0.097)
20% (10%)	Int	0.019 (0.094)	0.214 (0.041)	-0.083 (0.093)	0.470(0.064)	-0.460 (0.101)	0.867 (0.144)	0.115 (0.099)	0.501 (0.079)
	S.TE	-0.036 (0.188)	0.422 (0.131)	-0.193 (0.169)	0.688 (0.161)	-0.135 (0.193)	1.282 (0.277)	-0.173 (0.179)	0.586 (0.138)
	NS.TE	0.016 (0.137)	0.403 (0.099)	0.010 (0.115)	0.555 (0.100)	0.006 (0.142)	1.099 (0.193)	0.068 (0.121)	0.505 (0.088)

Table 22: Average Bias and MSE of 100 datasets, 100 samples by 225 metabolites for Bayes, zero, minimum and mean methods.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR. Simulation standard deviation in parentheses.

MNAR (MAR)		Bayes CI	Zero	Min	Mean
6%(3%)	**Power**	0.335 (0.108)	0.205 (0.065)	0.210 (0.065)	0.209 (0.062)
	Type1 Error	0.048 (0.034)	0.038 (0.019)	0.034 (0.016)	0.038 (0.019)
	AUC	0.770 (0.050)	0.692 (0.037)	0.709 (0.032)	0.698 (0.036)
10%(5%)	**Power**	0.329 (0.110)	0.192 (0.065)	0.177 (0.065)	0.198 (0.062)
	Type1 Error	0.050 (0.034)	0.039 (0.019)	0.032 (0.016)	0.040 (0.019)
	AUC	0.766 (0.050)	0.678 (0.037)	0.674 (0.032)	0.687 (0.036)
20%(10%)	**Power**	0.314 (0.108)	0.155 (0.065)	0.126 (0.065)	0.173 (0.062)
	Type1 Error	0.046 (0.034)	0.036 (0.019)	0.022 (0.016)	0.040 (0.019)
	AUC	0.759 (0.050)	0.649 (0.037)	0.625 (0.032)	0.668 (0.036)

Table 23: Average power, type 1 error and AUC of 100 datasets, 100 samples by 225 metabolites for Bayes, zero, minimum and mean methods.

Total missing was considered at 9%, 15% and 30%, and within each missing, MNAR was greater than MAR.

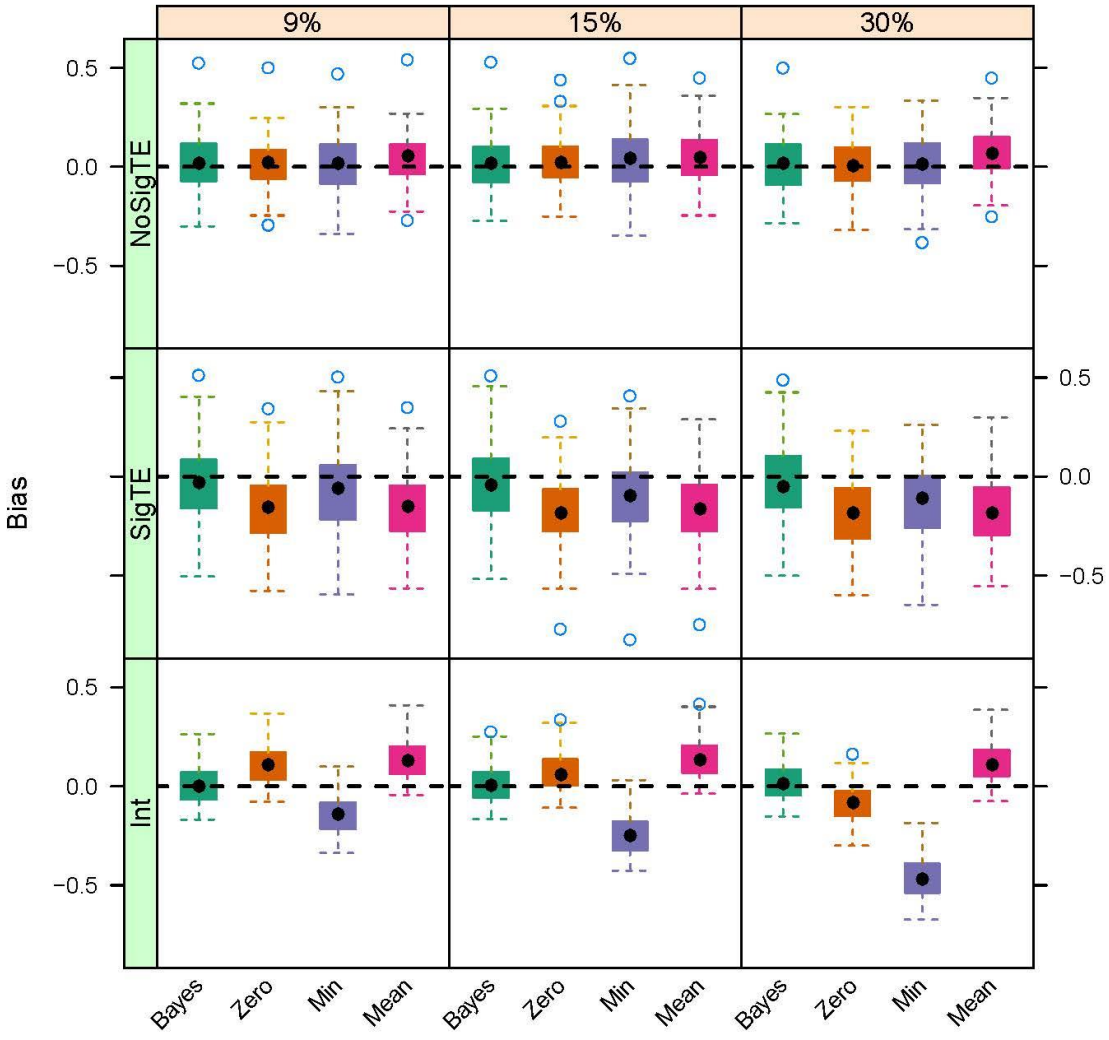

Figure 8: Boxplots of Bias for Bayes, Zero, Minimum and Means for 100 datasets, 10 samples by 225 metabolites.

Total missing was considered at 9%, 15% and 30% and within each missing MNAR is greater than MAR.

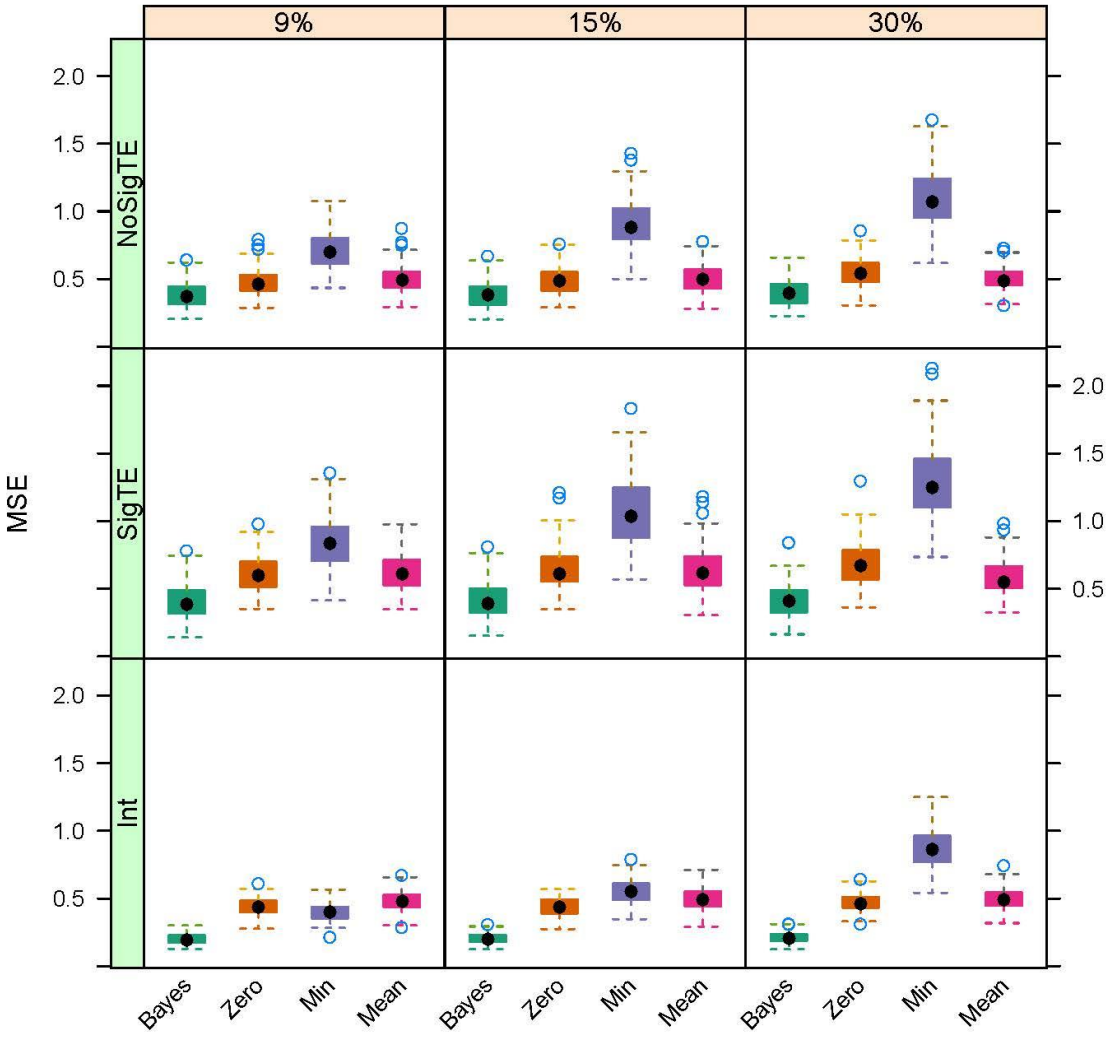

Figure 9: Boxplots of MSE for Bayes, Zero, Minimum and Means for 100 datasets, 10 samples by 225 metabolites.

Total missing was considered at 9%, 15% and 30% and within each missing MNAR is greater than MAR.

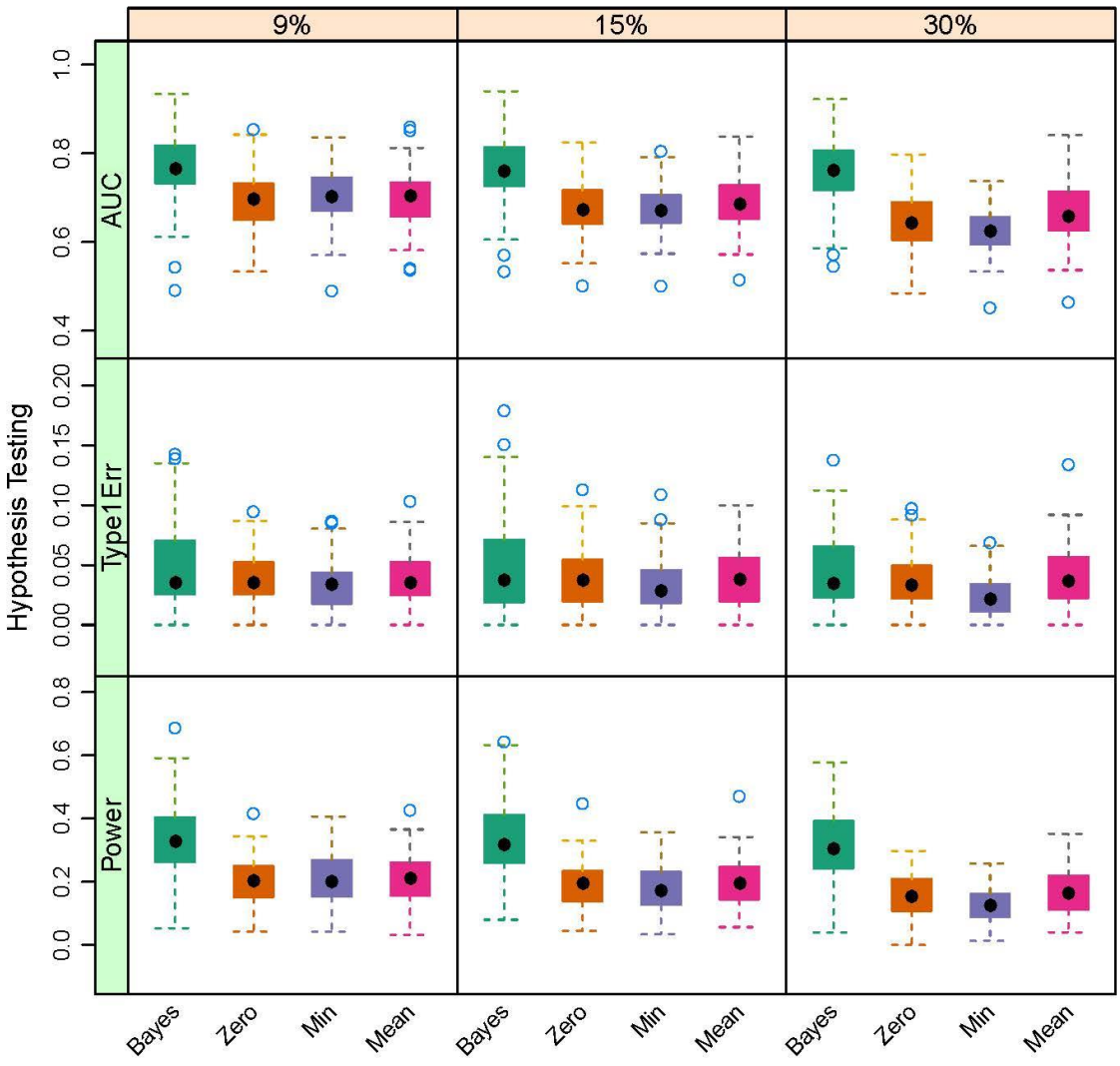

Figure 10: Boxplots of Power, Type1 Error and AUC for Bayes, Zero, Minimum and Means for 100 datasets, 10 samples by 225 metabolites.

Total missing was considered at 9%, 15% and 30% and within each missing MNAR is greater than MAR.

3.5 Discussion

In this work, we have proposed a Bayesian methodology to impute MVs, especially when the sample size is small. When metabolites occur at low abundance, below the detection limit of the instrumentation, we can consider it as missing not at random. In contrast, missing values resulting from technical errors are considered missing at random. To this end, we introduce a Bayesian model that incorporates the data augmentation that models imputing MVs handling the truncated data. Since MNAR is due to the detection limit, we consider the detection limit as a truncation point and assume that the metabolite follows a truncated normal distribution. The simulation results show that the proposed method performs better than standard approaches when there is missing data due to a threshold LOD.

In our simulations we evaluated only one size dataset which had small samples (10 samples by 225 metabolites). The LOD was calculated based on the missing percentage. For instance in 9% missing (where 6% was considered as MNAR) the 6% quantile for the complete data was considered as the LOD where we considered everything below that value as missing. The results shown in the tables are based on when the MNAR percentage is greater than the MAR percentage (e.g. for 9% total missing, 6% is MNAR and 3% is MAR). In the cleaning process (which removes metabolites with >50% MVs in each group) we are removing more metabolites whose values are concentrated near the LOD. For example in our case, after screening we reduced the metabolites to an average of about 207 metabolites for 15% missing and 186 metabolites for 30% missing out of the original 225. We included analysis looking at estimation efficiency (bias and MSE) and the impact on hypothesis testing whether metabolites have a significant treat effect or not.

An important limitation to our methodology is the reliance on the normality assumption for modeling the metabolite abundance. In our simulation study we investigated data from a normal distribution, whereas in many cases metabolite data may be non-normally distributed. In these cases we suggest to first transform the data to normality, then impute the values and lastly transform back. Hypothesis testing and inference may also be performed on transformation scale.

3.6 Conclusions

In conclusion, the simulation results reveal that compared with standard approaches such as zero, mean and min imputation, the Bayesian method is a coherent approach for imputing high dimensional data where there is missingness partially due to a truncation (detection) threshold. Results based on simulated data show that the Bayesian method generally has lower MSE values compared to the other three simpler imputation algorithms (zero, mean, and minimum value imputation) when there is both missing at random and missing due to a threshold value. Assessment based on hypothesis testing also demonstrates that the Bayesian method generally outperforms the other three methods. However, the approach has a computation expense due to the extensive iterations usage in MCMC algorithms. Even though this study is based on metabolomic datasets, our findings are more generally applicable to other types of high-dimensional data that contains missing values associated with an LOD, for instance proteomics data and delta-CT values from qRT-PCR array cards (Warner, Mukhopadhyay et al. 2014).

CHAPTER 4
BIOLOGICAL IMPACT OF IMPUTATION METHODS ON DOWNSTREAM ANALYSES

4.1 Background

High-throughput technologies, such microarrays or mass spectrometry (MS) suffer from missing values (MV) due to various experimental reasons. The issues posed by MVs are well-known in statistical data analysis literature (Little and Rubin, 2002). Standard downstream statistical methods have been developed to analyze complete data sets, where the rows represent the cases and the columns represent the variables measured. Although in many applications, these data matrices are not complete where variables for some cases cannot be measured technical problems, or the measurements are not reliable or obtainable for certain samples. Typically in high-throughput technology the missing values occur in a large number, e.g in metabolomics studies MVs are reported to comprise around 10–40% of data (Armitage et al, 2015) and thus it is not practical to simply remove samples with MVs as this can lead to selection bias.

Downstream analyses via multivariate methods require a complete dataset. MVs are handled differently and thus affects the interpretation and statistical inference. The most common approach used in handling MVs is case deletion, wherein this method only completed cases with no MVs are included in the analysis. This method leads to a smaller sample size which results in low power (White & Carlin, 2010; Harel et al., 2012). The other widely used approach is via single imputation methods where MVs are filled in with plausible values. It is a straight

forward method but also a dangerous way of dealing with missing values. Statistical analysis performed on datasets imputed by single imputation method may be biased as the approach does not consider the uncertainty of the imputed values. The common single imputation methods include mean, zero, half minimum and median imputation where the MVs are replaced by the mean, zeros, half of the minimum and median of the variable respectively. The magnitude of the covariances and correlation also decreases by limiting the variability, and this method often causes biased estimates, irrespective of the underlying missing data mechanism (Eekhout et al., 2012; Enders, 2010). Other single imputation methods that have been developed recently include the k-nearest neighbor (kNN), random forest (RF), Bayesian principal component analysis (BPCA), probabilistic principal component analysis (PPCA), and singular value decomposition (SVD) imputation (Schafer and Graham, 2002).

Within the microarray arena, there is noticeable presence of MV imputation methods and downstream analysis. In a recent comparative study by Brock et al (2008), eight MV imputation methods were investigated on different datasets and concluded that no single best MV imputation method exists but BPCA, LLS and LSA performed among the best. There is limited noticeable literature in the comparison of MV imputation methods in metabolomics. Gromski et al (2014), analyzed different MV imputation methods and their influence on multivariate analysis. They looked at five MV imputation methods: zero, mean, median, KNN, and Random Forest (RF) imputation and their influence on unsupervised and supervised learning and the final impact on the final output in terms of biological interpretation. Their results showed that the imputation methods have a considerable effect on the classification accuracy. Based on the data, they recommend that RF is better than the other methods as the classification rates for both supervised methods outperforms the other imputation methods. Another study by Hrydziuszko et al. (2011) summarized that the choice of imputation method

can significantly affect the results and interpretation of analyses of metabolomics data. They compared eight common MV imputation methods (substitution with a small predefined value, half minimum, mean, median, KNN, BPCA, multivariate imputation, REP (MV is substituted with average intensity of nearest peaks from the raw measurements of technical replicates)) and their impact on univariate and multivariate analysis. They concluded that the treatment of missing data is a very important step in data processing and suggest that KNN method imputes the most reliable values and is preferred method over the other methods. Armitage et al. (2015) tested four MV imputation methods (median, KNN, half minimum and zero) on different statistical tests and concluded that KNN was the best approximation for the real missing data.

There is noticeable absence in the literature of a comprehensive study of how the MV imputation methods affect the different downstream analyses in metabolomics such as biomarker detection, classification and cluster analysis. We perform a comprehensive and systematic evaluation to examine the biological impact of MV imputation in the three areas of downstream analyses: biomarker detection, classification and cluster analysis. To our knowledge, this is the first comprehensive evaluation study to focus on all three major downstream analyses.

4.2 Methods

To perform a comprehensive comparison and evaluation, we included six MV imputation methods and three major downstream analyses were considered; differential abundance, classification and clustering.

MV imputation methods

We included six MV imputation methods for evaluation: zero, minimum, mean, KNN-CR, KNN-TN, and KNN-EU.

Downstream analyses methods

We consider the three major types of downstream analyses to evaluate the biological impacts of MV imputation methods; differential abundance (DA) metabolite detection, classification and metabolite clustering. The specific methods evaluated are described below.

DA metabolite detection

We included the moderated t-test (limma), moderated t-test with fold-change (TREAT), and the standard t-test analysis. Smyth (2004) proposed an empirical Bayes (eBayes) approach which is a moderated version of the t-test that averages between the per-feature sample variance and a global (pooled) estimate of the variance. The TREAT method (McCartthy et al (2009)) is an extension to the eBayes method which tests whether differences in feature expression are above a given threshold.

Classification analysis

We included support vector machines (SVM), partial least squares discriminant analysis (PLS-DA) and k nearest neighbors (KNN).

Metabolite clustering analysis

We included *K*-means, hierarchical clustering and self-organizing maps (SOM) (Kohonen, 2001). Since the number of clusters *K* usually cannot be determined for a given dataset, we ran metabolite clustering using different choices of *K*, such as $K = 10$ and 15.

Assessment measures

We evaluated the performance of the imputation methods by using the root mean squared error (RMSE) as the metric on log transformed data. It measures the difference between the

estimated values and the original true values, when the original true values are known. The following simulation procedure from a complete dataset (CD) with no MVs is performed. MVs are generated by removing a proportion p of values from the complete data to generate data with MVs (MD). The MVs are then imputed as $\hat{y}_{im}$ (where y_{im} is the intensity of metabolite $m\ (1 \leq m \leq M)$ in sample $i\ (1 \leq i \leq N)$) using the given imputation method (ID). Finally, the root mean squared error (RMSE) is used to assess the performance by comparing the values of the imputed entries with the true values:

$$RMSE = \sqrt{\frac{1}{n(\mathcal{M})} \sum_{y_{im} \in \mathcal{M}} (\hat{y}_{im} - y_{im})^2},$$

where $\mathcal{M}$ is the set of missing values and $n(\mathcal{M})$ is the cardinality or number of elements in $\mathcal{M}$.

Metabolite list concordance index (MLCI) for DA metabolite detection

Suppose CD, MD and ID are obtained using the imputation methods and by applying a selected DA metabolite detection method (eBayes, TREAT, and t-test), we obtain a metabolite list from CD and another metabolite list from ID. The MLCI is defined as:

$$MLCI\ (M_{CD}, M_{ID}) = \frac{n(M_{CD} \cap M_{ID})}{n(M_{CD})} + \frac{n(M_{CD}^C \cap M_{ID}^C)}{n(M_{CD}^C)} - 1\,,$$

where M_{CD} is the list of statistically significant metabolites in the complete data, M_{ID} is the list of statistically significant metabolites in the imputed data, and M_{CD}^C and M_{ID}^C represent their complements, respectively. The metabolite list taken from the complete dataset is considered as the gold standard and a high value in MLCI indicates that the metabolite list from the imputed data is similar to that from the complete data. MLCI is correspondent to the Youden Index (Youden, 1950), which is defined as the sensitivity + specificity − 1.

Youden Index (YI)

We employ YI as a measure to evaluate the impact of MVs in classification. YI is defined as sensitivity + specificity − 1. We can directly calculate the YI of the prediction result from each imputed data because we know the true class labels of the samples in this supervised learning. We expect a good MV imputation method to generate a high YI.

Adjusted Rand Index (ARI)

The ARI (Hubert, 1985) is commonly used to evaluate the similarity between any two clustering results. Similar to MLCI we use the clustering result from the complete data as the gold standard and compare it with the imputed data clustering result. A higher ARI value indicates higher similarity between any two clustering results and that the MV imputation procedure leads to a smaller impact on the metabolite clustering analysis.

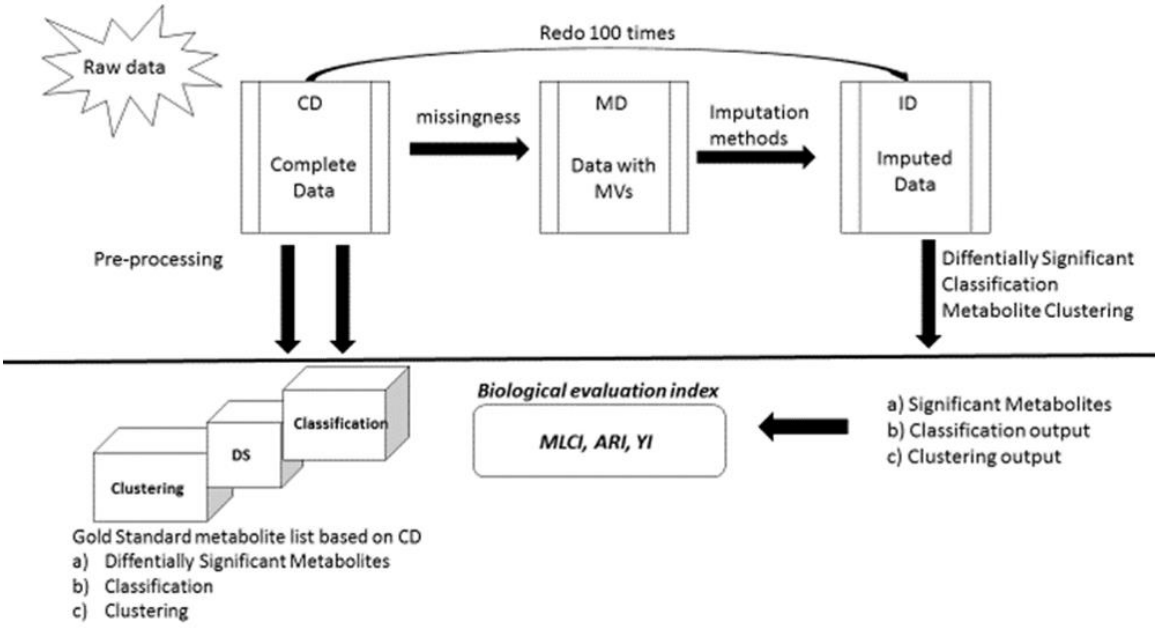

Figure 11: Schematic illustration of the research design

4.3 Real Data Studies

Atherothrombotic Data

We used the human atherothrombotic myocardial infarction (MI) metabolomics data. The data was identified between two groups, those with acute MI and those with stable coronary artery disease (CAD). Acute MI was further stratified into thrombotic (Type1) and non-thrombotic (Type2) MI. The data was collected across four time points and for the context of this research we used the baseline data only. The three groups, sCAD, Type1 and Type2 had 15, 11, and 12 patients with 1032 metabolites. The sCAD had 685 metabolites with complete values, 39 metabolites with complete missing, and 308 metabolites had 10.2% missing, the Type1 group had 689 metabolites with complete values, 43 metabolites with complete missing and 300 metabolites had 9.8% missing whereas the Type2 group had 610 metabolites with complete values, 66 metabolites with complete missing and 356 metabolites had 12.3% missing. The LOD for this dataset is considered as the minimum value of the dataset as commonly used in untargeted metabolomics. Plasma samples collected from the patients were used and 1032 metabolites were detected and quantified by GC-MS and ultra-performance (UP) LC-MS in both positive and negative ionization modes. Details of the experiment are described in DeFilippis et al (DeFilippis, Chernyavskiy et al. 2016).

Due to small sample sizes in metabolomics datasets, we used a simulation approach originally designed to resemble the multivariate distribution of gene expression in the original microarray data (Parrish, Spencer Iii et al. 2009). Since our Atherothrombotic data had missing values we first imputed missing values based on the KNN-CR method and then used the simulation method to simulate 100 datasets. The different groups were considered as independent datasets and the imputation was done on them separately. We used the similar mechanism for

missingness and screening as used in the simulation studies, with sample sizes of 25 and 50 for the human atherothrombotic dataset. For each simulation, we selected all the metabolites, and artificially generated a total of 60 differentially expressed metabolites with varying effect sizes between the two study groups (10 metabolites each with effect sizes 0.5, 1.0, 1.5, 2.0, 3.0, and 4.0). Sample sizes of 25 and 50 per group were evaluated, and a total of 100 simulations were conducted for each scenario. We used the generated simulated datasets for the three downstream analyses. For the DA, the p-Values from each method were adjusted based on the Benjamini–Hochberg (Benjamini Y et al (1995)) method to maintain an overall false-discovery rate of 0.05. Overall MLCI was estimated based on the mean of the 100 simulations. For the clustering analysis we clustered the metabolites based on 10 and 15 clusters.

4.4 Results

We conducted a simulation study based on the real datasets to further validate our results. Tables 24 (a-c) show the results of the MLCI based on the three DA metabolite detection test on the sCAD group of the human atherothrombotic data for the three different missing mechanism and two different sample sizes. Based on the table, we can see that KNN-TN performs little better than the rest of the imputation methods and the results seem consistent with the two sample sizes (25 and 50) and three missingness (9%, 15% and 30%). Table 25 (a-c) show the results of the ARI based on the three clustering methods on the sCAD group of the human atherothrombotic data for the three different missing mechanism and two different sample sizes. Table 26 (a-c) show the results of the YI based on the three classification methods on the sCAD group of the human atherothrombotic data for the three different missing mechanism and two different sample sizes. Figures 11, 12 and 13 plot the distribution

of the MLCI, ARI and YI for KNN-TN, KNN-CR, KNN-EU, Zero, Mean and Min by percent missing for the different methods used in each analyses for sample sizes 25.

Table 24: Average MLCI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU

SAMPLE SIZE	Test	Zero	Mean	Min	KNN-CR	KNN-TN	KNN-EU
25	**eBayes**	0.346 (0.081)	0.844 (0.065)	0.693 (0.087)	0.887 (0.051)	0.900 (0.048)	0.863 (0.060)
25	**t-test**	0.366 (0.083)	0.845 (0.063)	0.701 (0.082)	0.888 (0.050)	0.901 (0.046)	0.865 (0.059)
25	**TREAT**	0.433 (0.113)	0.731 (0.098)	0.683 (0.097)	0.817 (0.085)	0.832 (0.088)	0.756 (0.093)
50	**eBayes**	0.377 (0.078)	0.883 (0.052)	0.750 (0.069)	0.915 (0.045)	0.927 (0.041)	0.898 (0.051)
50	**t-test**	0.377 (0.078)	0.883 (0.053)	0.750 (0.071)	0.915 (0.046)	0.927 (0.041)	0.898 (0.051)
50	**TREAT**	0.492 (0.108)	0.768 (0.082)	0.748 (0.085)	0.853 (0.074)	0.870 (0.072)	0.787 (0.081)

Table 24a. Average MLCI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU. Total missing was considered at 9%.

SAMPLE SIZE	Test	Zero	Mean	Min	KNN-CR	KNN-TN	KNN-EU
25	**eBayes**	0.249 (0.067)	0.811 (0.071)	0.612 (0.080)	0.854 (0.060)	0.866 (0.058)	0.833 (0.066)
25	**t-test**	0.266 (0.068)	0.81 (0.070)	0.614 (0.074)	0.854 (0.059)	0.867 (0.056)	0.833 (0.065)
25	**TREAT**	0.363 (0.103)	0.665 (0.102)	0.585 (0.105)	0.772 (0.087)	0.793 (0.080)	0.697 (0.102)
50	**eBayes**	0.312 (0.077)	0.851 (0.054)	0.689 (0.075)	0.880 (0.045)	0.896 (0.042)	0.868 (0.050)
50	**t-test**	0.312 (0.076)	0.851 (0.052)	0.687 (0.074)	0.880 (0.044)	0.895 (0.043)	0.869 (0.049)
50	**TREAT**	0.443 (0.109)	0.700 (0.082)	0.659 (0.094)	0.807 (0.072)	0.833 (0.070)	0.724 (0.080)

Table 24b. Average MLCI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU. Total missing was considered at 15%.

SAMPLE SIZE	Test	Zero	Mean	Min	KNN-TN	KNN-CR	KNN-EU
25	**eBayes**	0.231 (0.063)	0.719 (0.075)	0.499 (0.083)	0.779 (0.067)	0.797 (0.065)	0.756 (0.070)
25	**t-test**	0.229 (0.061)	0.717 (0.075)	0.509 (0.087)	0.779 (0.069)	0.795 (0.067)	0.754 (0.075)
25	**TREAT**	0.359 (0.100)	0.485 (0.101)	0.423 (0.118)	0.671 (0.101)	0.689 (0.101)	0.540 (0.114)
50	**eBayes**	0.276 (0.066)	0.784 (0.066)	0.601 (0.077)	0.836 (0.051)	0.854 (0.050)	0.825 (0.056)
50	**t-test**	0.274 (0.066)	0.784 (0.066)	0.601 (0.078)	0.837 (0.051)	0.853 (0.050)	0.825 (0.057)
50	**TREAT**	0.417 (0.099)	0.551 (0.095)	0.516 (0.102)	0.733 (0.087)	0.756 (0.085)	0.598 (0.092)

Table 24c. Average MLCI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU. Total missing was considered at 30%.

Table 25: Average ARI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU

SAMPLE SIZE (K)	Test	Zero	Mean	Min	KNN-CR	KNN-TN	KNN-EU
25 (10)	**kMeans**	0.373 (0.032)	0.620 (0.086)	0.628 (0.088)	0.644 (0.097)	0.635 (0.105)	0.654 (0.113)
25 (10)	**hClust**	0.248 (0.053)	0.501 (0.073)	0.432 (0.082)	0.520 (0.078)	0.520 (0.080)	0.494 (0.079)
25 (10)	**SOM**	0.453 (0.012)	0.776 (0.029)	0.805 (0.019)	0.815 (0.029)	0.846 (0.031)	0.792 (0.029)
25 (15)	**kMeans**	0.290 (0.025)	0.597 (0.077)	0.571 (0.050)	0.624 (0.070)	0.629 (0.076)	0.604 (0.067)
25 (15)	**hClust**	0.224 (0.039)	0.475 (0.064)	0.399 (0.053)	0.495 (0.058)	0.511 (0.066)	0.475 (0.062)
25 (15)	**SOM**	0.421 (0.013)	0.728 (0.026)	0.757 (0.019)	0.773 (0.026)	0.805 (0.027)	0.748 (0.026)
50 (10)	**kMeans**	0.442 (0.028)	0.670 (0.082)	0.666 (0.083)	0.684 (0.096)	0.688 (0.106)	0.687 (0.096)
50 (10)	**hClust**	0.277 (0.065)	0.510 (0.079)	0.450 (0.066)	0.515 (0.078)	0.522 (0.084)	0.505 (0.078)
50 (10)	**SOM**	0.501 (0.011)	0.780 (0.023)	0.851 (0.019)	0.823 (0.024)	0.855 (0.024)	0.796 (0.024)
50 (15)	**kMeans**	0.365 (0.029)	0.613 (0.070)	0.599 (0.056)	0.623 (0.082)	0.617 (0.071)	0.622 (0.065)
50 (15)	**hClust**	0.261 (0.050)	0.495 (0.065)	0.435 (0.059)	0.496 (0.087)	0.510 (0.072)	0.484 (0.060)
50 (15)	**SOM**	0.469 (0.011)	0.731 (0.021)	0.792 (0.019)	0.786 (0.020)	0.818 (0.022)	0.750 (0.022)

Table 25a. Average ARI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU. K is the cluster size, K = 10 and 15 and the Sample size was 25 and 50 samples. Total missing was considered at 9%.

SAMPLE SIZE (K)	Test	Zero	Mean	Min	KNN-CR	KNN-TN	KNN-EU
25 (10)	**kMeans**	0.329 (0.022)	0.611 (0.078)	0.621 (0.075)	0.629 (0.102)	0.688 (0.106)	0.618 (0.087)
25 (10)	**hClust**	0.226 (0.041)	0.481 (0.068)	0.381 (0.089)	0.484 (0.076)	0.490 (0.073)	0.482 (0.074)
25 (10)	**SOM**	0.396 (0.013)	0.718 (0.022)	0.764 (0.019)	0.759 (0.025)	0.787 (0.027)	0.739 (0.023)
25 (15)	**kMeans**	0.256 (0.020)	0.531 (0.061)	0.520 (0.052)	0.577 (0.077)	0.577 (0.072)	0.552 (0.059)
25 (15)	**hClust**	0.206 (0.036)	0.454 (0.054)	0.341 (0.057)	0.460 (0.061)	0.466 (0.052)	0.452 (0.062)
25 (15)	**SOM**	0.373 (0.012)	0.666 (0.022)	0.702 (0.016)	0.698 (0.023)	0.722 (0.026)	0.687 (0.022)
50 (10)	**kMeans**	0.407 (0.026)	0.634 (0.084)	0.648 (0.080)	0.647 (0.090)	0.655 (0.100)	0.611 (0.081)
50 (10)	**hClust**	0.253 (0.063)	0.490 (0.071)	0.397 (0.076)	0.500 (0.087)	0.506 (0.079)	0.495 (0.079)
50 (10)	**SOM**	0.454 (0.011)	0.723 (0.022)	0.807 (0.020)	0.774 (0.023)	0.809 (0.031)	0.744 (0.022)
50 (15)	**kMeans**	0.360 (0.028)	0.531 (0.056)	0.534 (0.058)	0.577 (0.069)	0.591 (0.082)	0.533 (0.065)
50 (15)	**hClust**	0.239 (0.047)	0.463 (0.058)	0.357 (0.061)	0.474 (0.067)	0.485 (0.061)	0.466 (0.057)
50 (15)	**SOM**	0.422 (0.011)	0.664 (0.019)	0.729 (0.017)	0.705 (0.021)	0.736 (0.025)	0.684 (0.019)

Table 25b. Average ARI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU. K is the cluster size, K = 10 and 15 and the Sample size was 25 and 50 samples. Total missing was considered at 15%.

SAMPLE SIZE (K)	Test	Zero	Mean	Min	KNN-CR	KNN-TN	KNN-EU
25 (10)	**kMeans**	0.216 (0.017)	0.545 (0.058)	0.554 (0.060)	0.548 (0.065)	0.569 (0.069)	0.553 (0.058)
25 (10)	**hClust**	0.199 (0.044)	0.472 (0.092)	0.289 (0.094)	0.465 (0.083)	0.470 (0.080)	0.453 (0.090)
25 (10)	**SOM**	0.352 (0.014)	0.652 (0.020)	0.699 (0.020)	0.679 (0.023)	0.701 (0.036)	0.686 (0.020)
25 (15)	**kMeans**	0.154 (0.015)	0.517 (0.069)	0.448 (0.037)	0.525 (0.068)	0.535 (0.065)	0.528 (0.062)
25 (15)	**hClust**	0.169 (0.032)	0.443 (0.075)	0.266 (0.079)	0.434 (0.078)	0.455 (0.069)	0.436 (0.071)
25 (15)	**SOM**	0.338 (0.014)	0.583 (0.021)	0.634 (0.016)	0.611 (0.022)	0.638 (0.034)	0.618 (0.023)
50 (10)	**kMeans**	0.292 (0.019)	0.543 (0.062)	0.593 (0.063)	0.579 (0.077)	0.604 (0.079)	0.567 (0.059)
50 (10)	**hClust**	0.231 (0.052)	0.470 (0.079)	0.298 (0.087)	0.487 (0.0.85)	0.502 (0.079)	0.471 (0.087)
50 (10)	**SOM**	0.407 (0.012)	0.651 (0.017)	0.725 (0.020)	0.694 (0.019)	0.740 (0.003)	0.683 (0.018)
50 (15)	**kMeans**	0.377 (0.016)	0.883 (0.066)	0.750 (0.047)	0.915 (0.064)	0.927 (0.066)	0.898 (0.072)
50 (15)	**hClust**	0.377 (0.039)	0.883 (0.069)	0.750 (0.075)	0.915 (0.078)	0.927 (0.069)	0.898 (0.072)
50 (15)	**SOM**	0.492 (0.012)	0.768 (0.018)	0.748 (0.018)	0.853 (0.019)	0.870 (0.032)	0.787 (0.018)

Table 25c. Average ARI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU. K is the cluster size, K = 10 and 15 and the Sample size was 25 and 50 samples. Total missing was considered at 30%.

Table 26: Average YI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU.

SAMPLE SIZE	Test	Zero	Mean	Min	KNN-CR	KNN-TN	KNN-EU
25	**kNN**	0.214 (0.336)	0.607 (0.349)	0.607 (0.152)	0.643 (0.233)	0.732 (0.168)	0.643 (0.222)
25	**SVM**	0.875 (0.144)	0.964 (0.094)	0.982 (0.047)	1.000 (0.000)	1.000 (0.000)	0.964 (0.061)
25	**PLS-DA**	0.321 (0.313)	0.786 (0.213)	0.768 (0.112)	0.875 (0.144)	0.893 (0.112)	0.929 (0.098)
50	**kNN**	0.395 (0.237)	0.748 (0.172)	0.504 (0.151)	0.798 (0.95)	0.773 (0.161)	0.739 (0.159)
50	**SVM**	0.924 (0.065)	1.000 (0.000)	0.958 (0.044)	0.983 (0.044)	1.000 (0.000)	0.992 (0.022)
50	**PLS-DA**	0.605 (0.158)	0.983 (0.029)	0.933 (0.086)	0.992 (0.022)	0.992 (0.022)	0.983 (0.044)

Table 26a. Average YI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU. Total missing was considered at 9%.

SAMPLE SIZE	Test	Zero	Mean	Min	KNN-CR	KNN-TN	KNN-EU
25	**kNN**	0.312 (0.189)	0.734 (0.141)	0.375 (0.275)	0.703 (0.188)	0.781 (0.160)	0.562 (0.222)
25	**SVM**	0.766 (0.182)	0.984 (0.044)	0.953 (0.065)	0.969 (0.088)	1.000 (0.000)	0.969 (0.28)
25	**PLS-DA**	0.234 (0.279)	0.812 (0.211)	0.766 (0.156)	0.922 (0.133)	0.891 (0.141)	0.891 (0.141)
50	**kNN**	0.279 (0.188)	0.772 (0.192)	0.507 (0.206)	0.750 (0.160)	0.794 (0.104)	0.787 (0.109)
50	**SVM**	0.831 (0.228)	0.993 (0.021)	0.956 (0.052)	0.971 (0.044)	0.985 (0.042)	0.963 (0.062)
50	**PLS-DA**	0.566 (0.241)	0.971 (0.044)	0.956 (0.052)	0.993 (0.021)	0.978 (0.044)	0.971 (0.044)

Table 26b. Average YI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU. Total missing was considered at 15%.

SAMPLE SIZE	Test	Zero	Mean	Min	KNN-CR	KNN-TN	KNN-EU
25	**kNN**	0.050 (0.230)	0.612 (0.260)	0.388 (0.253)	0.725 (0.194)	0.788 (0.156)	0.700 (0.179)
25	**SVM**	0.625 (0.220)	0.900 (0.165)	0.875 (0.132)	0.975 (0.079)	1.000 (0.000)	0.962 (0.084)
25	**PLS-DA**	0.388 (0.190)	0.788 (0.196)	0.625 (0.177)	0.900 (0.129)	0.862 (0.138)	0.875 (0.156)
50	**kNN**	0.153 (0.222)	0.747 (0.194)	0.400 (0.179)	0.800 (0.206)	0.800 (0.136)	0.794 (0.147)
50	**SVM**	0.812 (0.067)	0.982 (0.028)	0.935 (0.070)	0.971 (0.064)	0.988 (0.025)	0.982 (0.040)
50	**PLS-DA**	0.347 (0.153)	0.976 (0.041)	0.865 (0.092)	0.994 (0.019)	0.994 (0.019)	0.976 (0.074)

Table 26c. Average YI of 100 simulations using the human atherothrombotic dataset sCAD group for Zero, Min, Means KNN-TN, KNN-CR and KNN-EU. Total missing was considered at 30%.

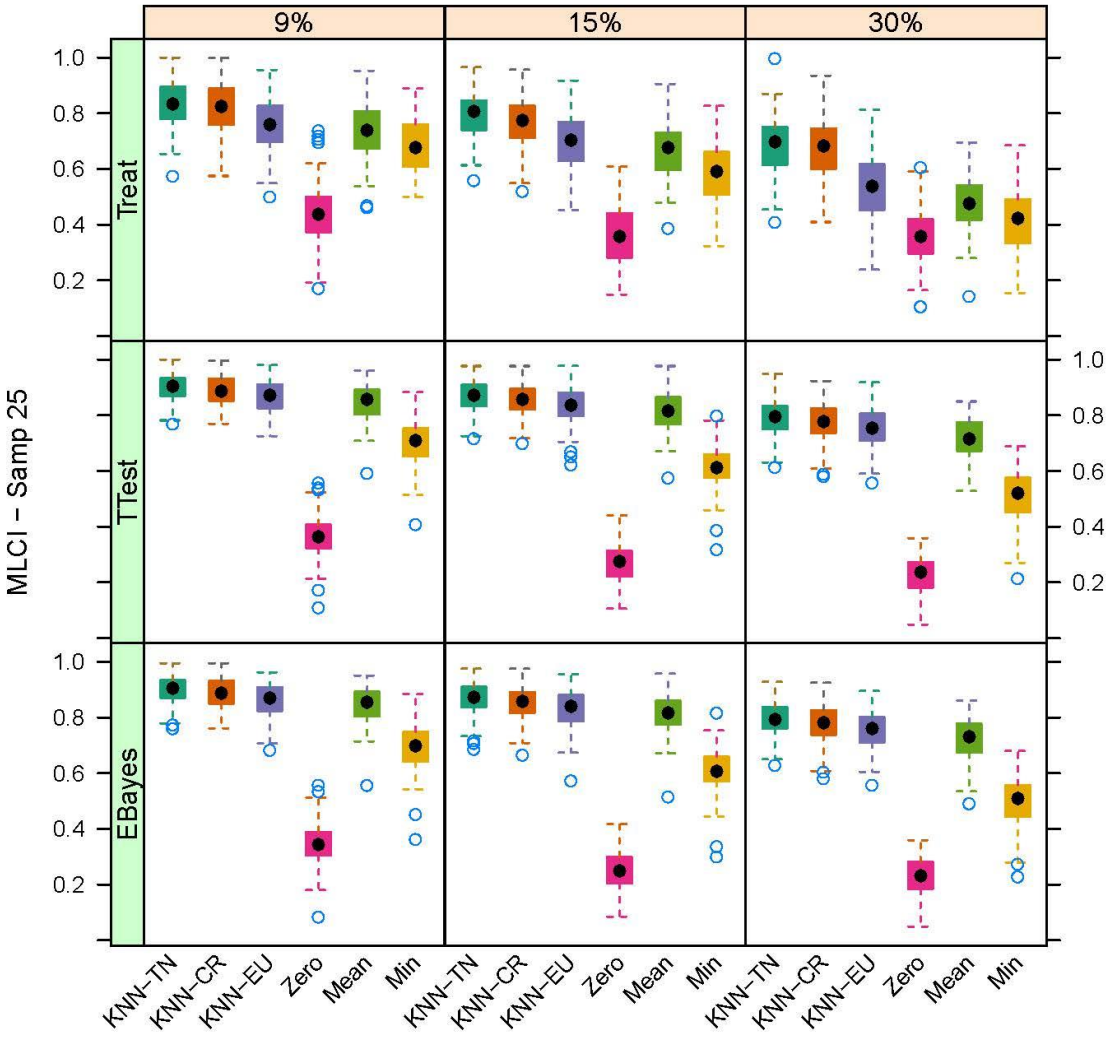

Figure 12: Boxplots of MLCI for KNN-TN, KNN-CR, KNN-EU, Zero, Mean and Min for 100 datasets, Sample size = 25.

Total missing was considered at 9%, 15% and 30% and within each missing MNAR is greater than MAR. The three differential abundance tests used were EBayes, TTest and TREAT.

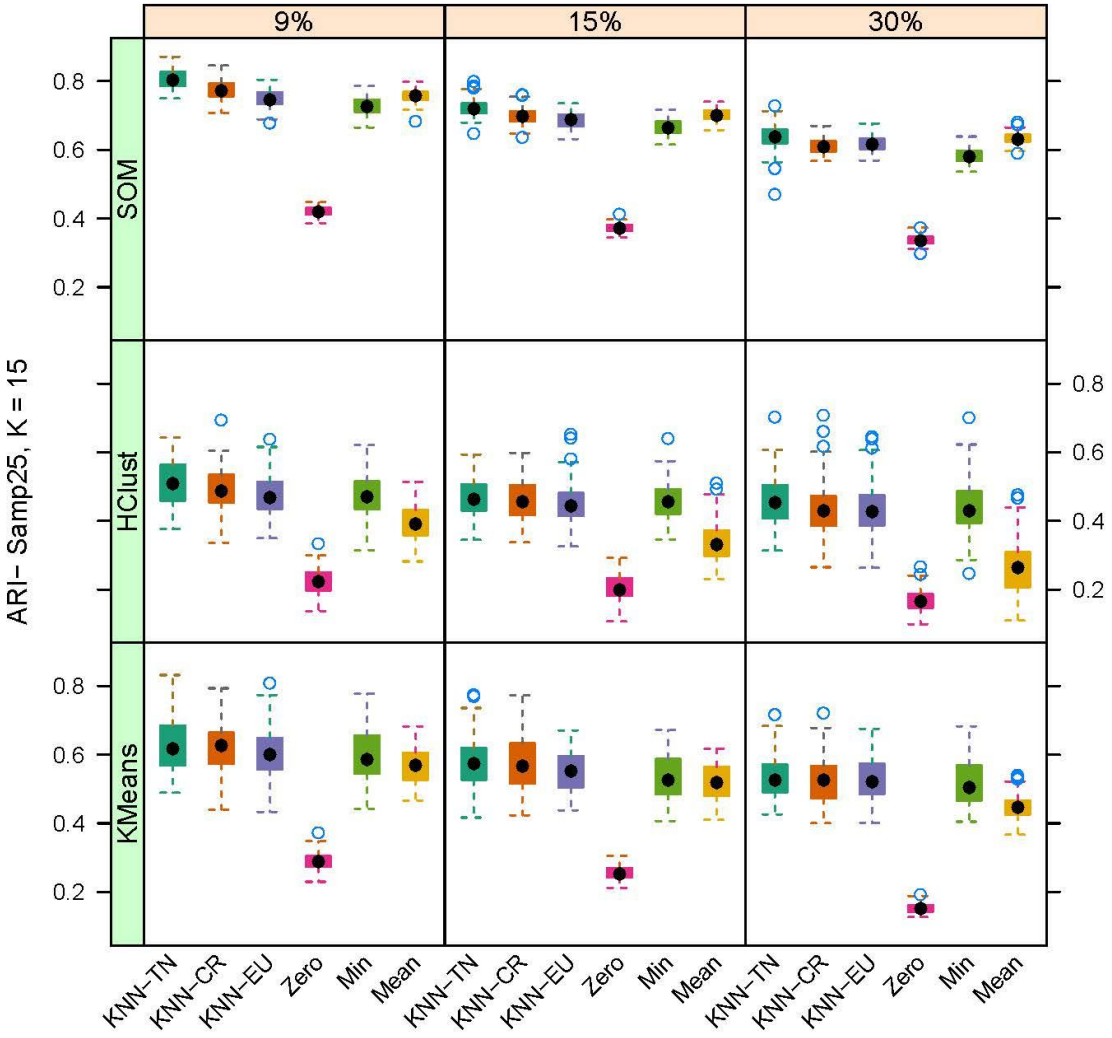

Figure 13: Boxplots of ARI for KNN-TN, KNN-CR, KNN-EU, Zero, Mean and Min for 100 datasets, Sample size = 25 and K = 15

Total missing was considered at 9%, 15% and 30% and within each missing MNAR is greater than MAR. The three clustering algorithm used were KMeans, HClust and SOM.

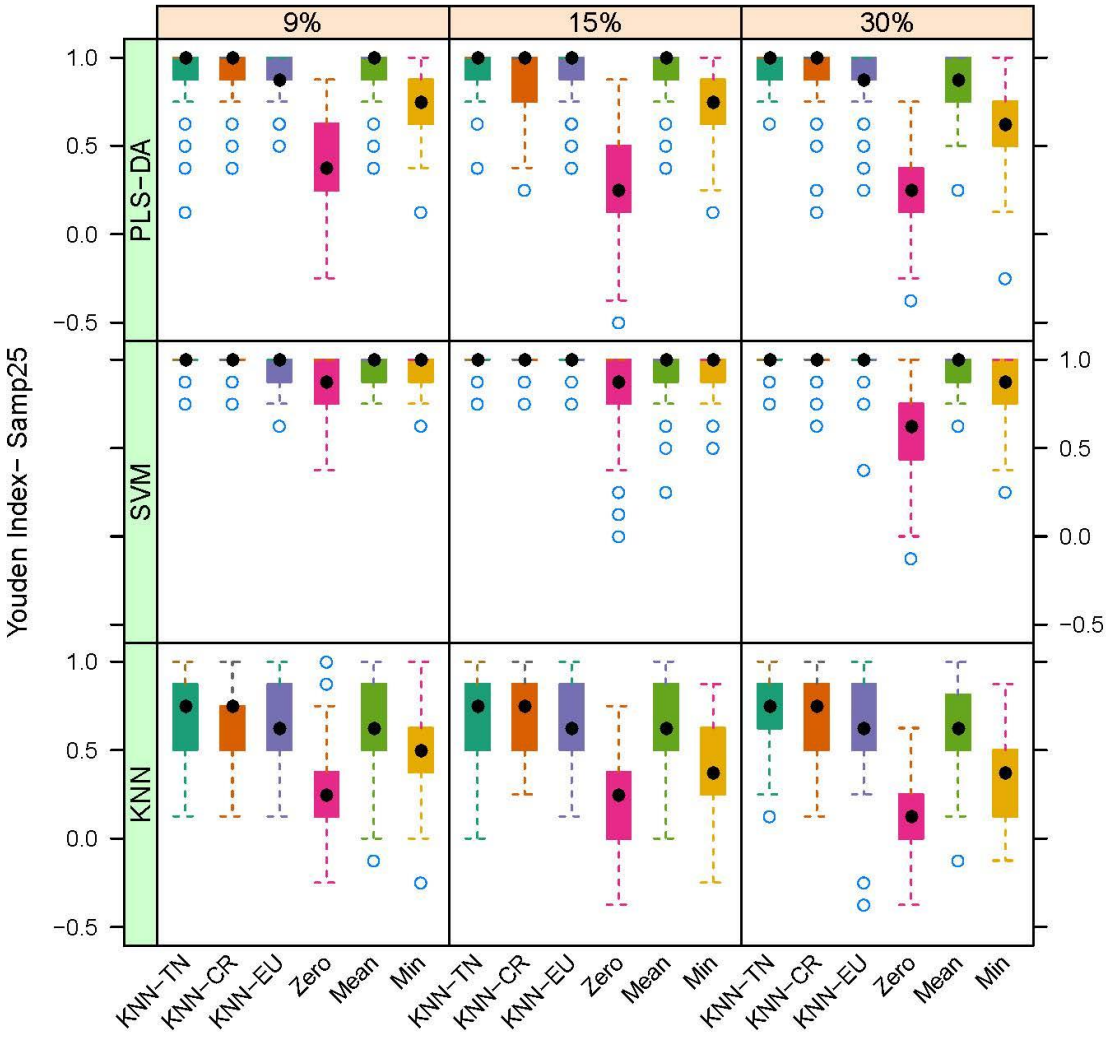

Figure 14: Boxplots of YI for KNN-TN, KNN-CR, KNN-EU, Zero, Mean and Min for 100 datasets, Sample size = 25

Total missing was considered at 9%, 15% and 30% and within each missing MNAR is greater than MAR. The three classification algorithms used were KNN, SVM and PLS-DA.

4.5 Discussions

The objective of this study was to evaluate the impact of MV imputation method on commonly performed downstream analyses. Although several prior studies have investigated the impact of MV imputation on individual analysis only, there has been no extensive analysis approach on the impact of MVs. Our investigation on the MVs imputation the RMSE measure was found to be the lowest with the KNN-TN method from the previous studies. In evaluating the biological impact of MV imputation on downstream analyses commonly carried out after MVs, we found that for detection of DA metabolites, the KNN-TN performed well compared to the other methods. In contrast for classification analysis, the impact of MV imputation was mixed. Overall, KNN-TN performed best in most cases, but KNN-CR or KNN-EU performed better in some cases. While DA metabolite detection and classification presented different results, the impact on clustering was also mixed where KNN-TN performed the best in most of the cases among the different methods. Our selection of biological impact measures for the effect of MV imputation on downstream analyses was motivated by choosing a measure that is both comprehensive and intuitive. The MLCI and YI were selected because they capture both the sensitivity and specificity of the result in a single measure, and the adjusted Rand index is a well-known and widely used measure of concordance between two clustering partitions. The primary analysis in our study is based on the downstream analysis of the logged data, a typical practice before analysis.

4.6 Conclusions

Based on the results, we conclude by highlighting the results from our study. Prior to deciding which imputation algorithm to use for MVs in metabolomics data, it is helpful for investigators to know which areas of downstream analysis are even impacted by MV imputation. The experimental results reveal that compared with KNN based on correlation and Euclidean

metrics, KNN based on truncation estimation is a competitive approach for all the three different downstream analyses. Results based on real data simulations show that the proposed method (KNN-TN) generally has higher MLCI, ARI and YI values compared to the other two KNN methods and simpler imputation algorithms (zero, mean, and minimum value imputation) when there is both missing at random and missing due to a threshold value.

CHAPTER 5
CONCLUSIONS AND FUTURE RESEARCH

This dissertation consisted of three research projects. The first project was a novel approach for MV imputation method based on truncated normal distribution with the nearest neighbors approach. The means and standard deviations of the metabolites were first estimated based on the truncated normal distribution with the LOD as the truncation point using the Newton Raphson algorithm. We conducted extensive simulation studies and real data set simulations to show that the proposed method outperforms from the standard approaches. With the parametric model, small sample size was a bottleneck and thus led to the second project where we employed data augmentation with a Bayesian model based on MCMC. The initial results on the Bayesian model show that with small samples the Bayesian model performs well compared to the other standard methods. We further evaluated the impact of MV imputation methods on three different downstream analyses; DA metabolite detection, classification and clustering. We conducted a comprehensive study to examine the impact on MV imputation methods and based on the results the KNN-TN method performs better or similar to KNN-CR but performs better compared to other methods. Detection of DA metabolites was the most sensitive analysis to the choice of imputation method while classification was the least sensitive and clusterting was intermediately affected.

The success of a MV imputation algorithm cannot be solely measured by its accuracy in imputing the underlying true values; its impact on downstream statistical inference is arguably

of greater importance. Further, the choice of an optimal imputation algorithm may depend on the underlying structure of the data. In the future studies, we want to extend and perform extensive simulations and extend the performance of MV imputation methods to other existing methods such as random forest, local least squares estimation etc. While previous studies have investigated the downstream impact of MV imputation in metabolomics, they have generally done so based on only a single imputation pass (Gromski et al and Hrydziuszko et al). By using multiple simulated datasets and more real datasets, we want to provide a more definite answer to the question of biological impact of missing value imputation in metabolomics. With the Bayesian framework, we want to extend the model to different correlation structures and evaluate more simulated data combinations and also apply on real datasets. We further want to specifically investigate how robust the Bayes method does by only using MNAR and comparing it with minimum imputation and then only using MAR and comparing it with the mean imputation. The impact of our novel approaches in project 1 and project 2 is limited without open-source software to implement them. We want to provide users with a robust package and user-friendly interface for missing value imputation in metabolomics studies. To permit multiple developers of the software, R packages will be hosted and developed on GitHub (https://github.com/) or RForge (https://r-forge.r-project.org/).

REFERENCES

"The Metabolomics WorkBench ", from http://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Project&ProjectID=PR000010.

Albrecht, D., Kniemeyer, O., Brakhage, A. A., & Guthke, R. (2010). Missing values in gel-based proteomics. *Proteomics*, *10*(6), 1202-1211.

Alonso, A., Marsal, S., & Julià, A. (2015). Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology*, *3*, 23.

Anders, S., & Huber, W. (2012). Differential expression of RNA-Seq data at the gene level–the DESeq package. *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)*.

Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, *78*(1), 40-64.

Armbruster, D. A., Tillman, M. D., & Hubbs, L. M. (1994). Limit of detection (LQD)/limit of quantitation (LOQ): comparison of the empirical and the statistical methods exemplified with GC-MS assays of abused drugs. *Clinical chemistry*, *40*(7), 1233-1238.

Armitage, E. G., Godzien, J., Alonso-Herranz, V., López-Gonzálvez, Á., & Barbas, C. (2015). Missing value imputation strategies for metabolomics data. *Electrophoresis*, *36*(24), 3050-3060.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Bothwell, J. H., & Griffin, J. L. (2011). An introduction to biological nuclear magnetic resonance spectroscopy. *Biological Reviews*, *86*(2), 493-510.

Brock, G. N., Shaffer, J. R., Blakesley, R. E., Lotz, M. J., & Tseng, G. C. (2008). Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC bioinformatics*, *9*(1), 12.

Clark, J. S. (2005). Why environmental scientists are becoming Bayesians. *Ecology letters*, *8*(1), 2-14.

Clark, J. S., & Gelfand, A. E. (2006). A future for models and data in environmental science. *Trends in Ecology & evolution*, *21*(7), 375-380.

Cohen Jr, A. C. (1949). On estimating the mean and standard deviation of truncated normal distributions. *Journal of the American Statistical Association*, *44*(248), 518-525.

Cohen Jr, A. C. (1950). Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. *The Annals of Mathematical Statistics*, 557-569.

Cole, R. F., Mills, G. A., Bakir, A., Townsend, I., Gravell, A., & Fones, G. R. (2016). A simple, low cost GC/MS method for the sub-nanogram per litre measurement of organotins in coastal water. *MethodsX*, *3*, 490-496.

DeFilippis, A. P., Chernyavskiy, I., Amraotkar, A. R., Trainor, P. J., Kothari, S., Ismail, I., Hargis, C. W., Korley, F. K., Leibundgut, G., Tsimikas, S., Rai, S. N., & Bhatnagar, A. (2016). Circulating levels of plasminogen and oxidized phospholipids bound to plasminogen distinguish between atherothrombotic and non-atherothrombotic myocardial infarction. *Journal of thrombosis and thrombolysis*, *42*(1), 61-76.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 327-351.

Depaoli, S., & van de Schoot, R. (2015). Improving Transparency and Replication in Bayesian Statistics: The WAMBS-Checklist.

Dunson, D. B. (2001). Commentary: practical advantages of Bayesian analysis of epidemiologic data. *American journal of Epidemiology*, *153*(12), 1222-1226.

Eekhout, I., de Vet, H. C., Twisk, J. W., Brand, J. P., de Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of clinical epidemiology*, *67*(3), 335-342.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical Association*, *72*(359), 557-565.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Fiehn, O. (2002). Metabolomics–the link between genotypes and phenotypes. *Plant molecular biology*, *48*(1-2), 155-171.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.

Gromski, P. S., Xu, Y., Kotze, H. L., Correa, E., Ellis, D. I., Armitage, E. G., Turner, M. L., & Goodacre, R. (2014). Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites*, *4*(2), 433-452.

Harel, O., Pellowski, J., & Kalichman, S. (2012). Are we missing the importance of missing values in HIV prevention randomized clinical trials? Review and recommendations. *AIDS and Behavior*, *16*(6), 1382-1393.

Hrydziuszko, O., & Viant, M. R. (2012). Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*, *8*(1), 161-174.

Karpievitch, Y., Stanley, J., Taverner, T., Huang, J., Adkins, J. N., Ansong, C., Heffron, F., Metz, T. O., Qian, W. J., Yoon, H., & Smith, R. D. (2009). A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics*, *25*(16), 2028-2034.

Karpievitch, Y. V., Dabney, A. R., & Smith, R. D. (2012). Normalization and missing value imputation for label-free LC-MS analysis. *BMC bioinformatics*, *13*(16), S5.

Katajamaa, M., & Orešič, M. (2007). Data processing for mass spectrometry-based metabolomics. *Journal of chromatography A*, *1158*(1), 318-328.

Kohonen, T. (2001). Self-organizing maps of massive databases. *International journal of engineering intelligent systems for electrical engineering and communications*, *9*(4), 179-186.

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, *39*(4), 653-686.

Lele, S. R., Dennis, B., & Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology letters*, *10*(7), 551-563.

Link, W. A., Cam, E., Nichols, J. D., & Cooch, E. G. (2002). Of BUGS and birds: Markov chain Monte Carlo for hierarchical modeling in wildlife research. *The Journal of wildlife management*, 277-291.

Little, R. J., & Rubin, D. B. (2002). Single imputation methods. *Statistical Analysis with Missing Data, Second Edition*, 59-74.

McCarthy, D. J., & Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, *25*(6), 765-771.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750-773.

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295-314.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 538-558.

Oh, S., Kang, D. D., Brock, G. N., & Tseng, G. C. (2011). Biological impact of missing-value imputation on downstream analyses of gene expression profiles. *Bioinformatics*, *27*(1), 78-86.

Oliver, S. G., Winson, M. K., Kell, D. B., & Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends in biotechnology*, *16*(9), 373-378.

Parrish, R. S., Spencer, H. J., & Xu, P. (2009). Distribution modeling and simulation of gene expression data. *Computational Statistics & Data Analysis*, *53*(5), 1650-1660.

Pedreschi, R., Hertog, M. L., Carpentier, S. C., Lammertyn, J., Robben, J., Noben, J. P., Panis, B., Swennen, R., & Nicolaï, B. M. (2008). Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. *Proteomics*, *8*(7), 1371-1383.

Ren, J. J., & Zhou, M. (2011). Full likelihood inferences in the Cox model: an empirical likelihood approach. *Annals of the Institute of Statistical Mathematics*, *63*(5), 1005-1018.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 581-592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Sadanala, K. C., Lee, J., Chung, B. C., & Choi, M. H. (2012). Targeted Metabolite Profiling: Sample Preparation Techniques for GC-MSBased Steroid Analysis. *Mass Spectrometry Letters*, *3*(1), 4-9.

Sansbury, B. E., De Martino, A. M., Xie, Z., Brooks, A. C., Brainard, R. E., Watson, L. J., DeFilippis, A. P., Cummins, T. D., Harbeson, M. A., Brittian, K. R., & Prabhu, S. D. (2014). Metabolomic analysis of pressure-overloaded and infarcted mouse hearts. *Circulation: Heart Failure*, CIRCHEARTFAILURE-114.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Schmidt, C. W. (2004). Metabolomics: what's happening downstream of DNA. *Environmental Health Perspectives*, *112*(7), A410.

Shah, J. S., Rai, S. N., DeFilippis, A. P., Hill, B. G., Bhatnagar, A., & Brock, G. N. (2017). Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. *BMC bioinformatics*, *18*(1), 114.

Shah, V. P., Midha, K. K., Findlay, J. W., Hill, H. M., Hulse, J. D., McGilveray, I. J., McKay, G., Miller, K. J., Patnaik, R. N., Powell, M. L., & Tonelli, A. (2000). Bioanalytical method validation—a revisit with a decade of progress. *Pharmaceutical research*, *17*(12), 1551-1557.

Shrivastava, A., & Gupta, V. B. (2011). Methods for the determination of limit of detection and limit of quantitation of the analytical methods. Chron Young Sci 2: 21–25.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, *3*(1), 3.

Smyth, G. K., Ritchie, M., Thorne, N., & Wettenhall, J. (2005). LIMMA: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., & Selbig, J. (2007). pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, *23*(9), 1164-1167.

Steuer, R., Morgenthal, K., Weckwerth, W., & Selbig, J. (2007). A gentle guide to the analysis of metabolomic data. *Metabolomics: Methods and protocols*, 105-126.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528-540.

Taylor, S. L., Leiserowitz, G. S., & Kim, K. (2013). Accounting for undetected compounds in statistical analyses of mass spectrometry 'omic studies. *Statistical applications in genetics and molecular biology*, *12*(6), 703-722.

Taylor, S. L., Ruhaak, L. R., Kelly, K., Weiss, R. H., & Kim, K. (2016). Effects of imputation on correlation: implications for analysis of mass spectrometry data from multiple biological matrices. *Briefings in bioinformatics*, bbw010.

Theodoridis, G., Gika, H. G., & Wilson, I. D. (2011). Mass spectrometry-based holistic analytical approaches for metabolite profiling in systems biology studies. *Mass spectrometry reviews*, *30*(5), 884-906.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botsein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520-525.

Tutz, G., & Ramzan, S. (2015). Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics & Data Analysis*, *90*, 84-99.

Wang, D., & Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, 490-517.

Want, E., & Masson, P. (2011). Processing and analysis of GC/LC-MS-based metabolomics data. *Metabolic Profiling: Methods and Protocols*, 277-298.

Warner, D. R., Mukhopadhyay, P., Brock, G., Webb, C. L., Michele Pisano, M., & Greene, R. M. (2014). MicroRNA expression profiling of the developing murine upper lip. *Development, growth & differentiation*, *56*(6), 434-447.

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine*, *29*(28), 2920-2931.

Xi, B., Gu, H., Baniasadi, H., & Raftery, D. (2014). Statistical analysis and modeling of mass spectrometry-based metabolomics data. *Mass spectrometry in metabolomics: methods and protocols*, 333-353.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32-35.

APPENDIX

Details of NR Procedure

For notational convenience we define the probability $P(Y \in (a, \infty)|\mu, \sigma^2)$ as

$$\zeta(\mu, \sigma^2) = \int_a^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy$$

Interchanging differentiation and integration, the first derivatives of the above equation with respect to μ and σ are

$$\zeta'_{\mu} = \int_a^{\infty} e^{\frac{-(y-\mu)^2}{2\sigma^2}} \times \frac{(y-\mu)}{\sigma^3\sqrt{2\pi}}\, dy, \qquad \text{and}$$

$$\zeta'_{\sigma} = \int_a^{\infty} e^{\frac{-(y-\mu)^2}{2\sigma^2}} \times \left(\frac{(y-\mu)^2}{\sigma^4\sqrt{2\pi}} - \frac{1}{\sigma^2\sqrt{2\pi}} \right) dy$$

Using the above derivatives, the gradient (G) (first partial derivative) with respect to the parameters is

$$\boldsymbol{G} = \begin{bmatrix} \frac{\partial l}{\partial \mu} \\ \frac{\partial l}{\partial \sigma} \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} -n\frac{\zeta'_{\mu}}{\zeta} - \frac{1}{\sigma^2}(n\mu - \sum y_i) \\ -n\frac{\zeta'_{\sigma}}{\zeta} - \frac{n}{\sigma} + \frac{\sum(y_i-\mu)^2}{\sigma^3} \end{bmatrix}$$

The second derivatives for the Hessian are

$$\zeta''_{\mu}(\mu,\sigma^2)=\frac{\partial^2\zeta}{\partial^2\mu}=\int_a^{\infty}e^{\frac{-(y-\mu)^2}{2\sigma^2}}\times\left(\frac{(y-\mu)^2}{\sigma^5\sqrt{2\pi}}-\frac{1}{\sigma^3\sqrt{2\pi}}\right)dy,$$

$$\zeta''_{\sigma}(\mu,\sigma^2)=\frac{\partial^2\zeta}{\partial^2\sigma}=\int_a^{\infty}e^{\frac{-(y-\mu)^2}{2\sigma^2}}\times\left(\frac{(y-\mu)^2}{\sigma^7\sqrt{2\pi}}-\frac{5(y-\mu)^2}{\sigma^5\sqrt{2\pi}}+\frac{2}{\sigma^3\sqrt{2\pi}}\right)dy,\qquad\text{and}$$

$$\psi''_{\mu,\sigma}(\mu,\sigma^2)=\frac{\partial^2\psi}{\partial\mu\partial\sigma}=\int_a^{\infty}e^{\frac{-(y-\mu)^2}{2\sigma^2}}\times\left(\frac{(y-\mu)^3}{\sigma^6\sqrt{2\pi}}-\frac{3(y-\mu)}{\sigma^4\sqrt{2\pi}}\right)dy$$

$$\zeta''_{\mu,\sigma}(\mu,\sigma^2)=\frac{\partial^2\zeta}{\partial\mu\partial\sigma}=\int_a^{\infty}e^{\frac{-(y-\mu)^2}{2\sigma^2}}\times\left(\frac{(y-\mu)^3}{\sigma^6\sqrt{2\pi}}-\frac{3(y-\mu)}{\sigma^4\sqrt{2\pi}}\right)dy$$

Using the equations above and taking the derivatives, the Hessian matrix is

$$\boldsymbol{H}=\begin{bmatrix}\frac{\partial g_1}{\partial\mu} & \frac{\partial g_1}{\partial\sigma}\\ \frac{\partial g_2}{\partial\mu} & \frac{\partial g_2}{\partial\sigma}\end{bmatrix}$$

$$=\begin{matrix}-n\frac{\zeta\zeta''_{\mu}-(\zeta'_{\mu})^2}{\zeta^2}-\frac{n}{\sigma^2} & -n\frac{\zeta\zeta''_{\sigma|\mu}-\zeta'_{\mu}\zeta'_{\sigma}}{\zeta^2}+\frac{2(n\mu-\sum y_i)}{\sigma^3}\\ -n\frac{\zeta\zeta''_{\mu|\sigma}-\zeta'_{\mu}\zeta'_{\sigma}}{\zeta^2}+\frac{2(n\mu-\sum y_i)}{\sigma^3} & -n\frac{\zeta\zeta''_{\sigma}-(\zeta'_{\sigma})^2}{\zeta^2}+\frac{n}{\sigma^2}-\frac{3(n\mu-\sum y_i)}{\sigma^4}\end{matrix}$$

CURRICULUM VITA

JASMIT S SHAH

Diabetes and Obesity Center
University of Louisville
Louisville, KY, 40202

Email: jasmit.shah@louisville.edu

EDUCATION:

2011-present	PhD Candidate (GPA 3.76) Department of Bioinformatics and Biostatistics University of Louisville Louisville, KY 40208
2009-2011	M.S. Bioinformatics and Biostatistics (GPA 3.82) Department of Bioinformatics and Biostatistics University of Louisville Louisville, KY 40208
2007-2009	B.S., Mathematics, Statistics and Chemistry (GPA 3.9) Department of Mathematics and Statistics University of South Alabama Mobile, AL 36688

PROFESSIONAL EXPERIENCES:

June 2014 – Current	Research Associate Organizing and conducting statistical data analysis of data generated in the Center. Assist in planning and developing research studies and grant proposals, and provide insight regarding their statistical validity. Diabetes and Obesity Center University of Louisville Louisville, KY 40202
June 2014 – Current	REDCap Database Coordinator Organizing and maintaining the REDCap Database for the

	projects conducted within the Center. Diabetes and Obesity Center University of Louisville Louisville, KY 40202
Jan 2014- June 2014	Data Analyst Comparison of Gene-Gene interactions and SNPs for Breast Cancer Survival. Evaluation of methods for analyzing gene-gene interaction data for survival outcomes. Examining differential expression of targeted gene sets in survival outcomes data. Department of Bioinformatics and Biostatistics University of Louisville Louisville, KY 40208
Aug 2011- Dec 2013	Graduate Research Assistant Bioinformatics Data Analysis. Developing Ensemble Regression method by combining different regression methods that perform better individually for proteomics data. Using the DAVID Gene Functional Classification tool to classify yeast and CAEEL gene list into functional related gene groups and also rank the importance of the discovered gene groups. Identification of protein with existing methods such as ProteinProphet and PeptideProphet. Identification of Post Translational Modifications from Tandem Mass Spectrometry (MS/MS) Data. University of Louisville Department of Biostatistics and Bioinformatics. Louisville, KY 40208
Jan 2011 – July 2011	Research Assistant Examining the importance of bottlenecks in protein networks. Evaluating different regression models with application to survival predictions using MALDI TOF MS dataset.
Jan 2010 – Dec 2010	Research Assistant Developing computational methods for analyzing DNA methylation data. Classification and variable importance of lung cancer cell lines. Different classification methods and how various variables were important was examined mainly on data with two types of lung cancer, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC).
Aug 2009 – Dec 2009	Data Analyst and Lab Assistant

	Bioinformatics Data Analysis. Statistical analysis of the differential expression of microRNA's in the brain and liver of the aging mice and rats. Gheens Center on Aging Research University of Louisville Louisville, KY 40208
May 2008 – May 2009	Math and Statistics Tutor University of South Alabama Department of Mathematics and Statistics Mobile, AL 36688

LEADERSHIP EXPERIENCES:

Jan 2013 – Current	Leader for Organization Assisting and Serving International Students (OASIS): RSO at University of Louisville
June 2011– Aug 2011	International Center Student Leader Helping International Literature Scholars for their Exchange Summer Program at University of Louisville
Sept 2010 – Apr 2011	Health Science Campus Liaison Facilitate student improvements between Belknap Campus and Health Science Campus at University of Louisville.

THESIS AND PUBLICATIONS:

Shah, J. S., An Adaptive Ensemble Learner Function via Bagging and Rank Aggregation with Applications to High Dimensional Data (Masters Thesis, August, 2011)

Shah, J. S., Rai, S. N., DeFilippis, A. P., Hill, B. G., Bhatnagar, A., and Brock, G. N. Distribution Based Nearest Neighbor Imputation for Truncated High Dimensional Data with Applications to Pre-Clinical and Clinical Metabolomics Studies. BMC Bioinformatics, 18(1):114.

Shah, J. S., Brock, G. N. and Rai, S. N. Metabolomics data analysis and missing value issues with application to infarcted mouse hearts. BMC Bioinformatics, 16(15):1-1 (2015)

DeJarnett, N., Yeager, R., Conklin, D., Lee, J., O'Toole, TE, McCracken, J., Abplanalp, W., Srivastava, S., Riggs, DW., Hamzeh, I., Wagner, S., Chugh, A., DeFilippis, A., Ciszewski, T., Wyatt, B., Becher, C., Higdon, D., Ramos, KS., Tollerud, DJ., Myers, JA., Rai, SN., **Shah, J. S.**, Zafar, N., Krishnasamy, SS., Prabhu, SD. and Bhatnagar, A. Residential Proximity to Major Roadways Is Associated With Increased Levels of AC133+ Circulating Angiogenic Cells. Arteriosclerosis, Thrombosis, and Vascular Biology, 10.1161/ATVBAHA.115.305724 (2015)

Cummins, T. D., Holden, C. R., Sansbury, B. E., Gibb, A. A., **Shah, J. S.**, Zafar, N., Tang, Y., Hellmann, J., Rai, S. N., Spite, M., Bhatnagar, A. and Hill, B. G. Metabolic remodeling of white adipose tissue in obesity. American Journal of Physiology- Endocrinology and Metabolism, 307, 3, E262-E277 (2014)

Shah, J. S., Datta, S. and Datta, S. A multi-loss super regression learner (MSLR) with application to survival prediction using proteomics. Computational Statistics, 29, 1749-1767 (2014)

Conklin, D. J., Malovichko, M. V., Zeller, I., Das, T. P., Krivokhizhina, T. V., Lynch, B. H., Lorkiewicz, P., Agarwal, A., Wickramasinghe, N., Haberzettl, P., Sithu, S. D., **Shah, J.**, O'Toole, T. E., Rai, S. N., Bhatnagar, A. and Srivastava, S. Biomarker of Exposure and Systemic Toxicity of Chronic Acrolein Inhalation in Mice. (Submitted Toxicological Sciences)

Malovichko, M. V., Zeller, I., Krivokhizhina, T. V., Zhengzhi, X., Lorkiewicz, P., Agarwal, A., Wickramasinghe, N., Sithu, S. D., **Shah, J.**, O'Toole, T. E., Rai, S. N., Bhatnagar, A. Conklin, D. J. and Srivastava, S. Systemic Toxicity of Smokeless Tobacco Products in mice. (Submitted Nicotine and Tobacco Research)

Shah, J. S., Rai, S. N., Bhatnagar, A., Brock, G. N and Gaskins, J. Bayesian Modeling for Missing Value Imputation with Applications to Metabolomics. (In Print)

Shah, J. S., Rai, S. N., Bhatnagar, A., and Brock, G. N. Biological Impact of Missing Value Imputation on Down-stream Analyses of Metabolomic Profiles. (In Print)

Shah, J. S., Rai, S. N., Bhatnagar, A., and Brock, G. N. Statistical Methods and Analysis of Metabolomics Datasets (Review) (In Print)

Shah, J. S., Rai, S. N., Bhatnagar, A., and Brock, G. N. How to use Statistics on Flow Cytometry Data with R (In Print)

Shah, J. S., Zhang J., Witliff, J., Andres, S., Kidd La C., and Brock G. N. Comparison of Approaches for Detecting Gene-Gene Interactions for Survival Data (In Print)

POSTERS

Shah, J. S., Brock, G. N., Bhatnagar, A. and Rai, S. N. The Impact and Influence of Missing Value Imputation on Down-stream Analyses of Metabolomic Profiles. Research Louisville 2016, Louisville, KY, Oct 11-14 2016. **(First Place)**

Shah, J. S., Brock, G. N., Bhatnagar, A. and Rai, S. N. Truncation-Based Nearest Neighbors Imputation for High Dimensional Data with Detection Limit Thresholds. ENAR 2016, Austin, TX, March 6- 9 2016.

Shah, J. S., Brock, G. N., Bhatnagar, A. and Rai, S. N. Truncation-Based Nearest Neighbors Imputation for High Dimensional Data with Detection Limit Thresholds. Research Louisville 2015, Louisville, KY, Oct 27-30 2015. **(First Place)**

Shah, J. S., Brock, G. N. and Rai, S. N. Issues in the Statistical Analysis in Metabolomics Data with Application to Pressure-Overloaded and Infarcted Mouse Hearts. JSM 2015, Seattle, WA, Aug 8 -13 2015.

Shah, J. S., Brock, G. N. and Rai, S. N. Metabolomics Data Analysis and Missing Value Issues With Application to Infarcted Mouse Hearts. UT-ORNL-KBRIN Bioinformatics Summit 2015, Paris Landing State Park Buchanan, TN, March 20 -22 2015.

Zafar, N., Krishnasamy, S. S., **Shah, J. S.**, McCracken, J., Holden, C. R., DeJarnett, N. K., Abplanalp, W., Hill, B., Conklin, D., O'Toole, T., Rai, S. and Bhatnagar, A., Depletion of Circulating CD34+/KDR+ Cells in Type 2 Diabetes is Associated With Glycemic Control. AHA Scientific Sessions, Chicago, IL, November 15 – 19 2014

Zafar, N., Bhatnagar, A., Krishnasamy, S., O'Toole, T., Rai, S. N., **Shah, J. S.**, McCracken, J., Abplanalp, W., Hill, B. and Conklin, D. Depletion of circulating CD34+/KDR+ cells in Type 2 Diabetes is associated with glycemic control. Research Louisville 2014, Louisville KY, September 16 – 18 2014

Shah, J.S., Datta, S. Identification of Post Translational Modifications from Tandem Mass Spectrometry (MS/MS) Data. SRCOS 2013 Poster Competition, Burns TN, June 2-5 2013.

Dwyer, A., Mansoor, T., Nayak, V., **Shah, J.S.**, Datta, S. and Brier, M. Vascular Access and Daily Hemodialysis: A Clinical Experience. Poster. ASDIN 2013 Scientific Meeting, Washington, DC, February 15-17 2013.

WORSHOPS/MEETINGS/CONFERENCES

AHA- Tobacco Regulation and Addiction Center Annual Meeting, Louisville, KY, March 2017

ENAR 2017, Washington DC, March 2017

Fostering Diversity in Biostatistics Workshop ENAR 2017

Research Louisville 2016, Louisville, KY, October 2016

Joint Statistical Meetings 2016, Chicago, IL, August 2016

Fostering Diversity in Biostatistics Workshop JSM 2016

AHA- Tobacco Regulation and Addiction Center Annual Meeting, Louisville, KY, April 2016

UT-ORNL-KBRIN Bioinformatics Summit 2016, Lake Barkley State Park. KY, April 2016

ENAR 2016, Austin, TX, March 2016

Fostering Diversity in Biostatistics Workshop ENAR 2016

Research Louisville 2015, Louisville, KY, October 2015

Joint Statistical Meetings 2015, Seattle, WA, August 2015

UAB 3rd Metabolomics Workshop 2015, Birmingham, AL, June 2015

UT-ORNL-KBRIN Bioinformatics Summit 2015, Paris Landing State Park, TN, March 2015

AHA- Tobacco Regulation and Addiction Center Annual Meeting, Louisville, KY, March 2015

KBRIN Next Generation Sequencing (NGS) Workshop, Lexington, KY, July 2014

UT-ORNL-KBRIN Bioinformatics Summit 2014, Lake Barkley State Park. KY, April 2014

Southern Regional Council on Statistics (SRCOS), Burns, TN, June 2013

UT-ORNL-KBRIN Bioinformatics Summit 2013, Paris Landing State Park, TN, March 2013

SCHOLARSHIPS:

Travel Award SRCOS 2013 Burns TN, June 2-5 2013

KBRIN Support Scholarship, University of Louisville, 2010 - 2011

Scholarship Award, University of Louisville International Center, 2010

Scholarship Award, University of Louisville International Center, 2009

Scholarship Award, University of South Alabama Department of Mathematics and Statistics, 2009

Scholarship Award, University of South Alabama Department of Mathematics and Statistics, 2008

Sushila Mishra Memorial Mathematics and Statistics Scholarship, University of South Alabama, Spring 2009

TECHNICAL SKILLS:

Statistics: Microarray Data Analysis, Bioinformatics Data Analysis, SAM analysis

Statistical Software: R, SAS, SPSS, Minitab, WinBUGS, Matlab, Bioconductor, Sigma plot

Other application software: Microsoft Office, LaTeX, NimbleScan software, DAVID Clustering, Sample Size Calculators, Octave, Mathematica

Biological Applications: BLAST, NCBI, KEGG pathway, UniProt, Swiss-Prot, TrEMBL, GenBank, Ensembl Genome Browser, UCSC Genome Browser, EPD, TRASFAC, Protein Information Resource, dbSNP, miRbase, miRanda and Target Scan

Programming Languages: C++, Java, Perl, SQL

PROFESSIONAL SOCIETIES:

American Statistical Association (since 2010)

The International Biometrics Society Eastern North American Region (since 2016)

Rotaract Club of Greater Louisville (Secretary: Sept 2015 – Dec 2016)

Student Chapter ASA University of Louisville (Treasurer: Aug 2015 – Aug 2016)

HONORS:

Deans Citation Award 2017 (Nominated and Awarded)

University of Louisville Dean's List Aug 2009 – Dec 2013

Golden Key International Honour Society 2011- Current

Summa Cum Laude May 2a009

Best Math and Statistics Student May 2009

University of South Alabama Deans List Jan 2008 – May 2009

Macomb Community College Deans List Jan 2004 – May 2007

Top Student in Mathematics Kenya 2002

Silver Standard of the President's Award Kenya: 2002

Bronze Standard of the President's Award Kenya: 2001

LANGUAGE SKILLS:

Gujarati (Mother tongue)

English (Read, write and speak)

Hindi (Read, write and speak)

Swahili (Read, write and speak)

VOLUNTEERING ACTIVITIES

Aug 2011 – Dec 2016: International OASIS University of Louisville (Head Leader)

Jan 2016 – Current: Gujarati Samaj of Greater Louisville (Nonprofit 501c3) (Board)

Aug 2015 – Aug 2016: Biostatistics Club University of Louisville (Treasurer)